Universality and the circular law for sparse random matrices

Philip Matchett Wood Department of Mathematics
Stanford University
Building 380
Stanford, CA 94305
pmwood@math.stanford.edu

October 11, 2010

Abstract

The universality phenomenon asserts that the distribution of the eigenvalues of random matrix with iid zero mean, unit variance entries does not depend on the underlying structure of the random entries. For example, a plot of the eigenvalues of a random sign matrix, where each entry is $+1$ or -1 with equal probability, looks the same as an analogous plot of the eigenvalues of a random matrix where each entry is complex Gaussian with zero mean and unit variance. In the current paper, we prove a universality result for sparse random n by n matrices where each entry is non-zero with probability $1/n^{1-\alpha}$ where $0 < \alpha \leq 1$ is any constant. The sparse universality result proves convergence in probability and has one additional hypothesis that the real and imaginary parts of the entries are independent (this hypothesis is most likely an artifact of the proof). One consequence of the sparse universality principle is that the circular law holds for sparse real random matrices so long as the entries have zero mean and unit variance, which is the most general result for sparse real matrices to date.

1 Introduction

Given an n by n complex matrix A, we define the *empirical spectral distribution* (which we will abbreviate *ESD*), to be the following discrete probability measure on $\mathbb{C}$:

$$\mu_A(z) := \frac{1}{n} \left|\{1 \leq i \leq n : \mathrm{Re}(\lambda_i) \leq \mathrm{Re}(z) \text{ and } \mathrm{Im}(\lambda_i) \leq \mathrm{Im}(z)\}\right|,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A with multiplicity. In this paper, we focus on the case where A is chosen from a probability distribution on $\mathbb{M}_n(\mathbb{C})$, the set of all n by n complex matrices, and thus μ_A is a randomly generated discrete probability measure on $\mathbb{C}$.

1.1 Background: universality and the circular law

Suppose that A_n is an n by n matrix with iid random entries, each having zero mean and unit variance. The distribution of the eigenvalues of $(1/\sqrt{n})A_n$ approaches the uniform distribution on

Mathematics Subject Classification: 15A52
Keywords: random matrces, sparse, circular law, eigenvalues

the unit disk as n goes to infinity, a phenomenon known as the circular law. The non-sparse circular law has been proven in many special cases by many authors, including Mehta [15] (Gaussian case), Girko [7, 8], Edelman [6] (real Gaussian case), Bai [1] and Bai-Silverstein [2] (continuous case with bounded $(2+\delta)$th moment, for $\delta > 0$), Götze-Tikhomirov [9] (sub-Gaussian case) and [10] (bounded $(2+\delta)$th moment, for $\delta > 0$), Pan-Zhao [17] (bounded 4th moment), and Tao-Vu [25] (bounded $(2+\delta)$th moment, for $\delta > 0$). The following, due to Tao and Vu [27, Theorem 1.10], is the current best result, requiring only zero mean and unit variance.

Theorem 1.1 (Non-sparse circular law)**.** [27, Theorem 1.10] *Let X_n be the n by n random matrix whose entries are iid complex random variables with mean zero and variance one. Then the ESD of $\frac{1}{\sqrt{n}}X_n$ converges (both in probability and in the almost sure sense) to the uniform distribution on the unit disk.*

Proving convergence in the almost sure sense is in general harder than proving convergence in probability, and in the current paper, we will focus exclusively on convergence in probability. See Subsection 1.4 towards the end of the introduction for a description of convergence in probability and in the almost sure sense for the current context.

In [27], Tao and Vu ask the following natural question: what analog of Theorem 1.1 is possible in the case where the matrix is sparse, where entries become more likely to be zero as n increases, instead of entries having the same distribution for all n? One goal of the current paper is to provide an answer to this question in the form of Theorem 1.6 (see below), which proves the circular law for sparse random matrices with iid entries with the additional assumption that the real and complex parts of each entry are independent. In Figure 1, parts (b) and (d) give examples of the non-sparse circular law for Bernoulli and Gaussian random variables, and parts (a) and (c) give examples of the sparse circular law for Bernoulli and Gaussian random variables.

The literature studying the eigenvalues of sparse random matrices is distinctly smaller than that for non-sparse random matrices. Most authors have focused on studying the eigenvalues in the symmetric case, including [20, 18, 5, 21, 16, 19, 13, 22]. There has been, however, some recent and notable progress for non-symmetric sparse random matrices. Götze and Tikhomirov [9, 10] provide sparse versions for their proofs of the circular law with some extra conditions. In [9] they use the additional assumptions that the entries are sub-Gaussian and that each entry is zero with probability ρ_n where $\rho_n n^4 \to \infty$ as $n \to \infty$, and in [10] they use the additional assumption that the entries have bounded $(2+\delta)$th moment. The strongest result in the literature for non-symmetric sparse random matrices is due to Tao and Vu [25] who in 2008 proved a sparse version of the circular law with the assumption of bounded $(2+\delta)$th moment (note that [25] proves almost sure convergence, rather than convergence in probability as shown by [9, 10]).

Theorem 1.2. [25, Theorem 1.3] *Let $\alpha > 0$ and $\delta > 0$ be arbitrary positive constants. Assume that x is a complex random variable with zero mean and finite $(2+\delta)$th moment. Set $\rho = n^{-1+\alpha}$ and let A_n be the matrix with each entry an iid copy of $\frac{1}{\sqrt{\rho}}\mathbb{I}_\rho x$, where $\mathbb{I}_\rho$ is a random variable independent of x taking the value 1 with probability ρ and the value 0 with probability $1-\rho$. Let $\mu_{\frac{1}{\sigma\sqrt{n}}A_n}$ be the ESD of $\frac{1}{\sigma\sqrt{n}}A_n$, where σ^2 is, as usual, the variance of x. Then $\mu_{\frac{1}{\sqrt{\sigma n}}A_n}$ converges in the almost sure sense to the uniform distribution μ_∞ over the unit disk as n tends to infinity.*

In this paper, we prove a sparse circular law without the bounded $(2+\delta)$th moment condition, with our work being motivated by the proof in [27] of the (non-sparse) circular law in the general zero mean, unit variance case.

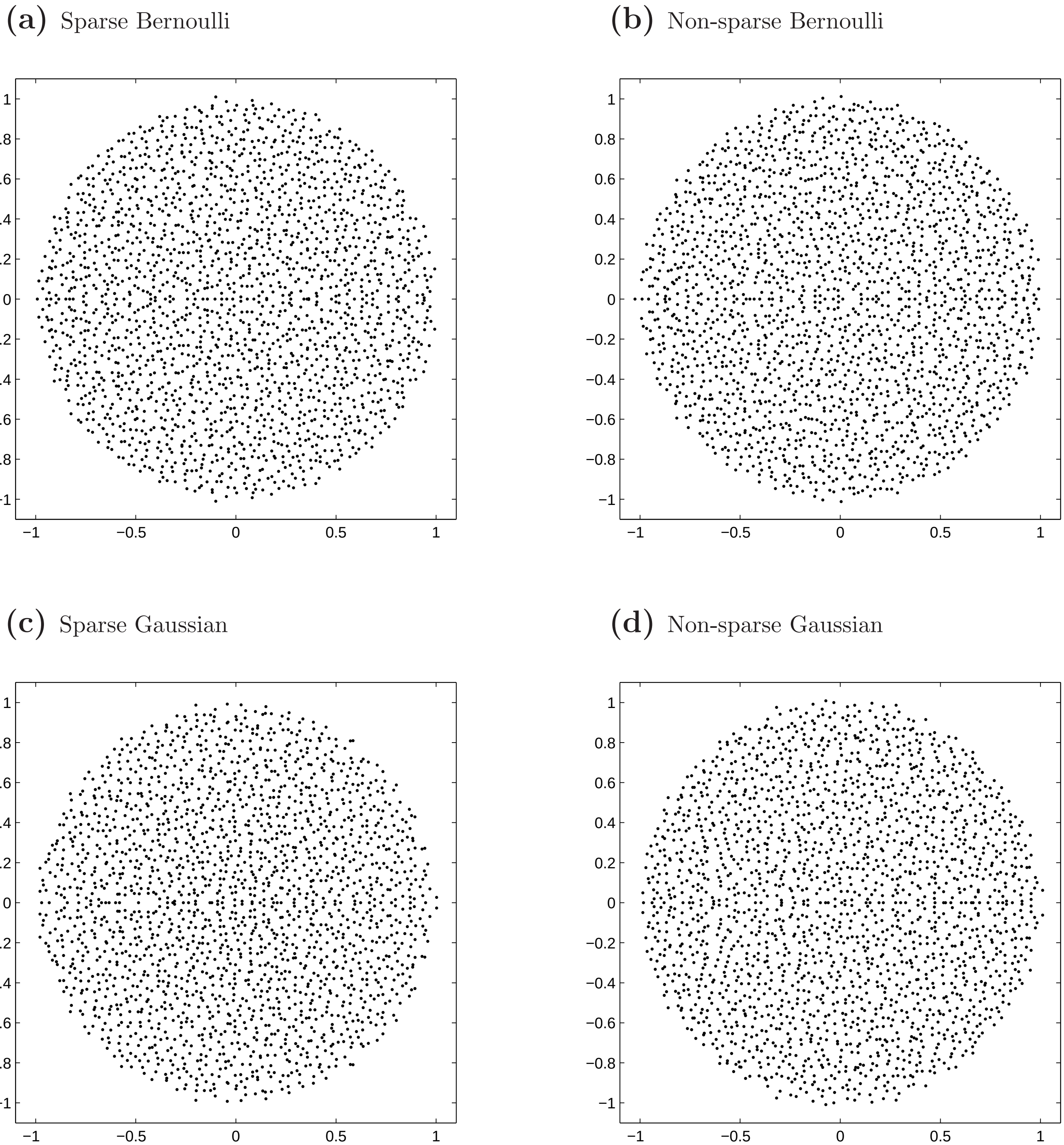


Figure 1: The four figures above illustrate that the circular law holds for for Bernoulli and Gaussian random matrix ensembles in both the sparse and non-sparse cases. Each plot is of the eigenvalues of a 2000 by 2000 random matrix with iid entries. In the first column (figures (a) and (c)), the matrices are sparse with parameter $\alpha = 0.4$, which means each entry is zero with probability $1 - \frac{1}{n^{0.6}}$, and in the second column (figures (b) and (d)), the matrices are not sparse (i.e., $\alpha = 1$). In the first row, both matrix ensembles are Bernoulli, so each non-zero entry is equally likely to be -1 or 1, and in the second row, the ensembles are Gaussian, so the non-zero entries are drawn from a Gaussian distribution with mean zero and variance one.

There has been much recent interest in demonstrating universal behavior for the eigenvalues of various types of random matrices. The following theorem is a fundamental result from [27]. For a matrix $A = (a_{ij})_{1\le i,j\le n}$, we will use $\|A\|_2$ to denote the Hilbert-Schmidt norm, which is defined by $\|A\|_2 = \operatorname{trace} AA^* = \left(\sum_{1\le i,j\le n} |a_{ij}|^2\right)^{1/2}$.

Theorem 1.3 (Universality principle)**.** [27] *Let x and y be complex random variables with zero mean and unit variance. Let $X_n := (x_{ij})_{1\le i,j\le n}$ and $Y_n := (y_{ij})_{1\le i,j\le n}$ be $n \times n$ random matrices whose entries x_{ij}, y_{ij} are iid copies of x and y, respectively. For each n, let M_n be a deterministic $n \times n$ matrix satisfying*

$$\sup_n \frac{1}{n^2} \|M_n\|_2^2 < \infty. \tag{1}$$

Let $A_n := M_n + X_n$ and $B_n := M_n + Y_n$. Then $\mu_{\frac{1}{\sqrt{n}}A_n} - \mu_{\frac{1}{\sqrt{n}}B_n}$ converges in probability to zero.

The universality principle as proven in [27, Theorem 1.5] also includes an additional hypothesis under which $\mu_{\frac{1}{\sqrt{n}}A_n} - \mu_{\frac{1}{\sqrt{n}}B_n}$ converges almost surely to zero (see [27] for details). In [27], Tao and Vu suggest the project of extending their universality principle for random matrices to the case of sparse random matrices. In this paper, we will follow the program developed in [27] and prove a universality principle for sparse random matrices.

1.2 New results for sparse random matrices

We begin by defining the type of sparse matrix ensemble that we will consider in this paper.

Definition 1.4 (Sparse matrix ensemble)**.** Let $0 < \alpha \le 1$ be a constant, and let $\mathbb{I}_\rho$ be the random variable taking the value 1 with probability $\rho := n^{-1+\alpha}$ and the value 0 with probability $1 - \rho$. Let x be a complex random variable that is independent of $\mathbb{I}_\rho$. The *n by n sparse matrix ensemble for x with parameter α* is defined to be the matrix X_n where each entry is an iid copy of $\frac{1}{\sqrt{\rho}}\mathbb{I}_\rho x$.

The main result of the current paper is the following:

Theorem 1.5 (Sparse universality principle)**.** *Let $0 < \alpha \le 1$ be a constant, and let x be a random variable with mean zero and variance one. Assume that the real and complex parts of x are independent; namely, that $\operatorname{Re}(x)$ is independent of $\operatorname{Im}(x)$. Let X_n be the n by n sparse matrix ensemble for x with parameter α, and let Y_n be the n by n matrix having iid copies of x for each entry (in particular, Y_n is* not *sparse). For each n, let M_n be a deterministic n by n matrix such that*

$$\sup_n \frac{1}{n^2} \|M_n\|_2^2 < \infty, \tag{2}$$

and let $A_n := M_n + X_n$ and $B_n := M_n + Y_n$. Then, $\mu_{\frac{1}{\sqrt{n}}A_n} - \mu_{\frac{1}{\sqrt{n}}B_n}$ converges in probability to zero.

Figure 2 gives an illustration of Theorem 1.5 with non-trivial M_n for sparse and non-sparse Bernoulli and Gaussian ensembles. In [12] a description is given for the asymptotic distribution of the ESD of a random matrix of the form $M_n + X_n$, where M_n is an arbitrary diagonal matrix.

Relating the sparse case to the non-sparse case in the above theorem is quite useful, since many results are known for random matrices with non-sparse iid entries, including a number of results in [27]. One of the motivating consequences of Theorem 1.5 is the following result, which is a combination of Theorem 1.5 and Theorem 1.1, the non-sparse circular law proven in [27].

Theorem 1.6 (Sparse circular law)**.** *Let $0 < \alpha \leq 1$ be a constant and let x be a random complex variable with mean zero and variance one, such that* $\mathrm{Re}(x)$ *and* $\mathrm{Im}(x)$ *are independent. Let X_n be the sparse matrix ensemble for x with parameter α. Then the ESD for $\frac{1}{\sqrt{n}}X_n$ converges in probability to the uniform distribution on the unit disk.*

An illustration of Theorem 1.6 appears in Figure 1. Note that the sparse circular law (Theorem 1.6) does not hold when $\alpha = 0$, since the probability of a row of all zeroes approaches a constant as $n \to \infty$, and thus with probability tending to 1 as $n \to \infty$, a constant fraction of the rows contain all zeroes.

In the non-sparse case, [27] also gives a number of extensions and generalizations, one of which is the circular law for shifted matrices, including the case where the entries of a random matrix have constant, non-zero mean.

Theorem 1.7 (Non-sparse circular law for shifted matrices)**.** [27, Corollary 1.12] *Let X_n be the n by n random matrix whose entries are iid complex random variables with mean 0 and variance 1, and let M_n be a deterministic matrix with rank $o(n)$ and obeying Inequality* (1)*. Let $A_n := M_n + X_n$. Then the ESD of $\frac{1}{\sqrt{n}}A_n$ converges (both in probability and in the almost sure sense) to the uniform distribution on the unit disk.*

Because Theorem 1.7 applies to non-sparse matrices of the form $M_n + X_n$, it can be directly combined with the sparse universality principle of Theorem 1.5 to yield the following result:

Theorem 1.8 (Sparse circular law for shifted matrices)**.** *Let $0 < \alpha \leq 1$ be a constant, and let x be a complex random variable with mean 0 and variance 1 such that the real and complex parts of x are independent (i.e.,* $\mathrm{Re}(x)$ *and* $\mathrm{Im}(x)$ *are independent). Let X_n be the n by n sparse random matrix ensemble with parameter α, let M_n be a deterministic matrix with rank $o(n)$ and obeying Inequality* (1)*, and let $A_n := M_n + X_n$. Then the ESD of $\frac{1}{\sqrt{n}}A_n$ converges in probability to the uniform distribution on the unit disk.*

An example of Theorems 1.7 and 1.8 appears in Figure 3.

The simple lemma below is a critical component for adapting arguments from [27] to the sparse case and illustrates a critical transition that occurs when $\alpha = 0$.

Lemma 1.9. *Let ξ be a complex random variable such that $\mathbb{E}\,|\xi| < \infty$. Let X be a sparse version of ξ, namely $X := \mathbb{I}_\rho \xi/\rho$, where $\rho = n^{-1+\alpha}$, where $0 < \alpha \leq 1$ is a constant. Then*

$$\mathbb{E}\Big(\Big|\mathbb{1}_{\{X > n^{1-\alpha/2}\}} X\Big|\Big) \to 0,$$

as $n \to \infty$.

Proof. The key steps to this proof are using independence of $\mathbb{I}_\rho$ and ξ, and applying monotone convergence. We compute:

$$\begin{aligned}\mathbb{E}\Big(\Big|\mathbb{1}_{\{|X| > n^{1-\alpha/2}\}} X\Big|\Big) &= \mathbb{E}\Big(\Big|\mathbb{1}_{\{|\mathbb{I}_\rho \xi| > n^{\alpha/2}\}} \mathbb{I}_\rho \xi/\rho\Big|\Big) \\ &\leq \frac{1}{\rho}\mathbb{E}\Big(\Big|\mathbb{1}_{\{|\xi| > n^{\alpha/2}\}} \mathbb{I}_\rho \xi\Big|\Big) \\ &= \mathbb{E}\Big(\Big|\mathbb{1}_{\{|\xi| > n^{\alpha/2}\}} \xi\Big|\Big).\end{aligned}$$

Finally, $\mathbb{E}\Big(\Big|\mathbb{1}_{\{\xi > n^{\alpha/2}\}} \xi\Big|\Big) \to 0$ as $n \to \infty$ by monotone convergence. $\square$

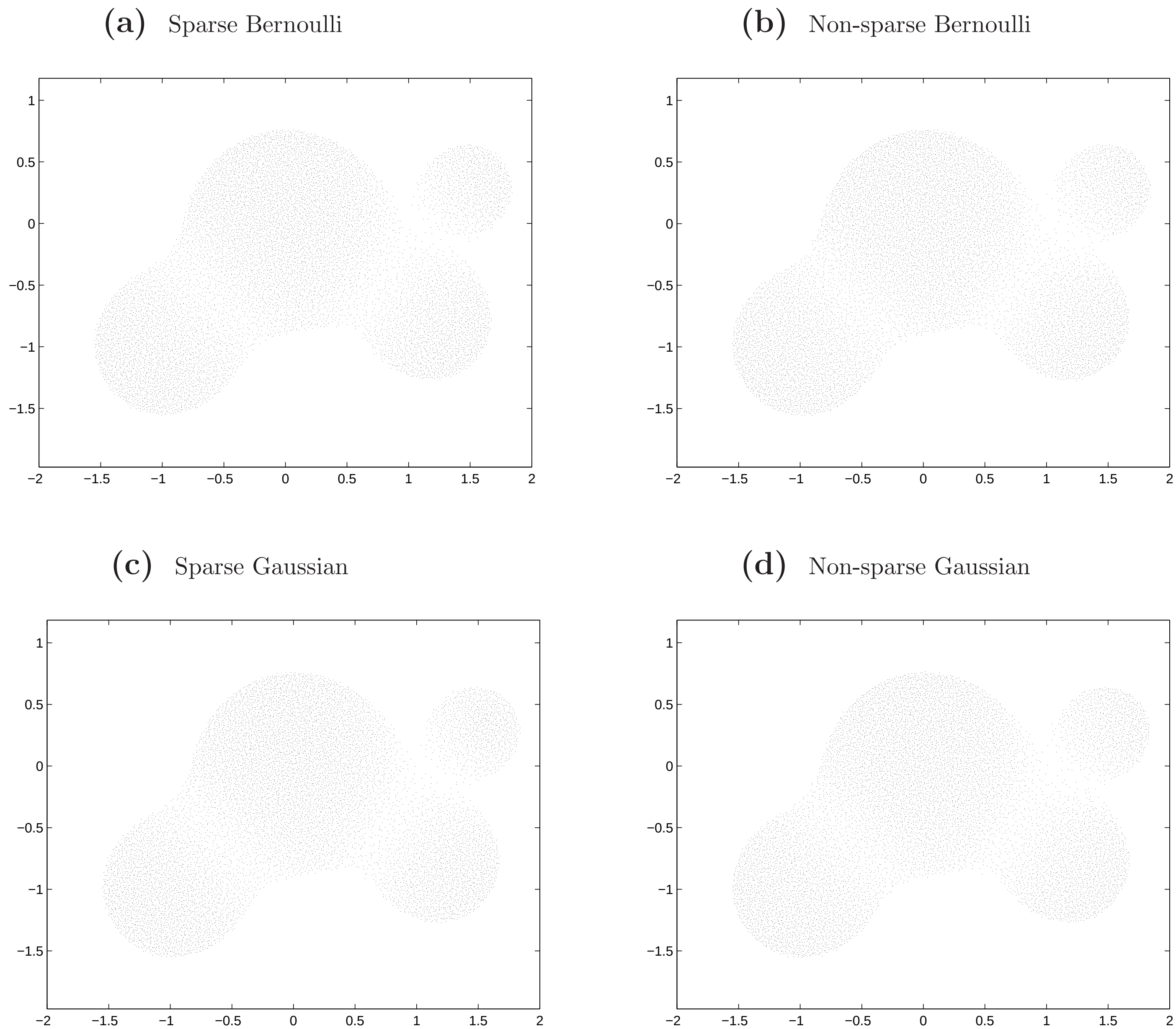


Figure 2: The four plots above illustrate that the universality principle holds for Bernoulli and Gaussian random matrix ensembles in both the sparse and non-sparse cases. Each plot is of the eigenvalues of a $10,000$ by $10,000$ random matrix with of the form $M_n + X_n$, where M_n is a fixed, non-random matrix and X_n contains iid entries. For each of the four plots, $\frac{1}{\sqrt{n}} M_n$ is the diagonal matrix with the first $\lfloor n/4 \rfloor$ diagonal entries equal to $-1-\sqrt{-1}$, the next $\lfloor n/6 \rfloor$ diagonal entries equal to $1.2 - 0.8\sqrt{-1}$, the next $n/12$ diagonal entries equal to $1.5 + 0.3\sqrt{-1}$, and the remaining entries equal to zero. In the first column (figures (a) and (c)), the matrices X_n are sparse with parameter $\alpha = 0.5$, which means each entry is zero with probability $1 - \frac{1}{n^{0.5}}$, and in the second column (figures (b) and (d)), the matrices X_n are not sparse (i.e., $\alpha = 1$). In the first row, both matrix ensembles are Bernoulli, so each non-zero entry of X_n is equally likely to be -1 or 1, and in the second row, the ensembles are Gaussian, so the non-zero entries of X_n are drawn from a Gaussian distribution with mean zero and variance one.

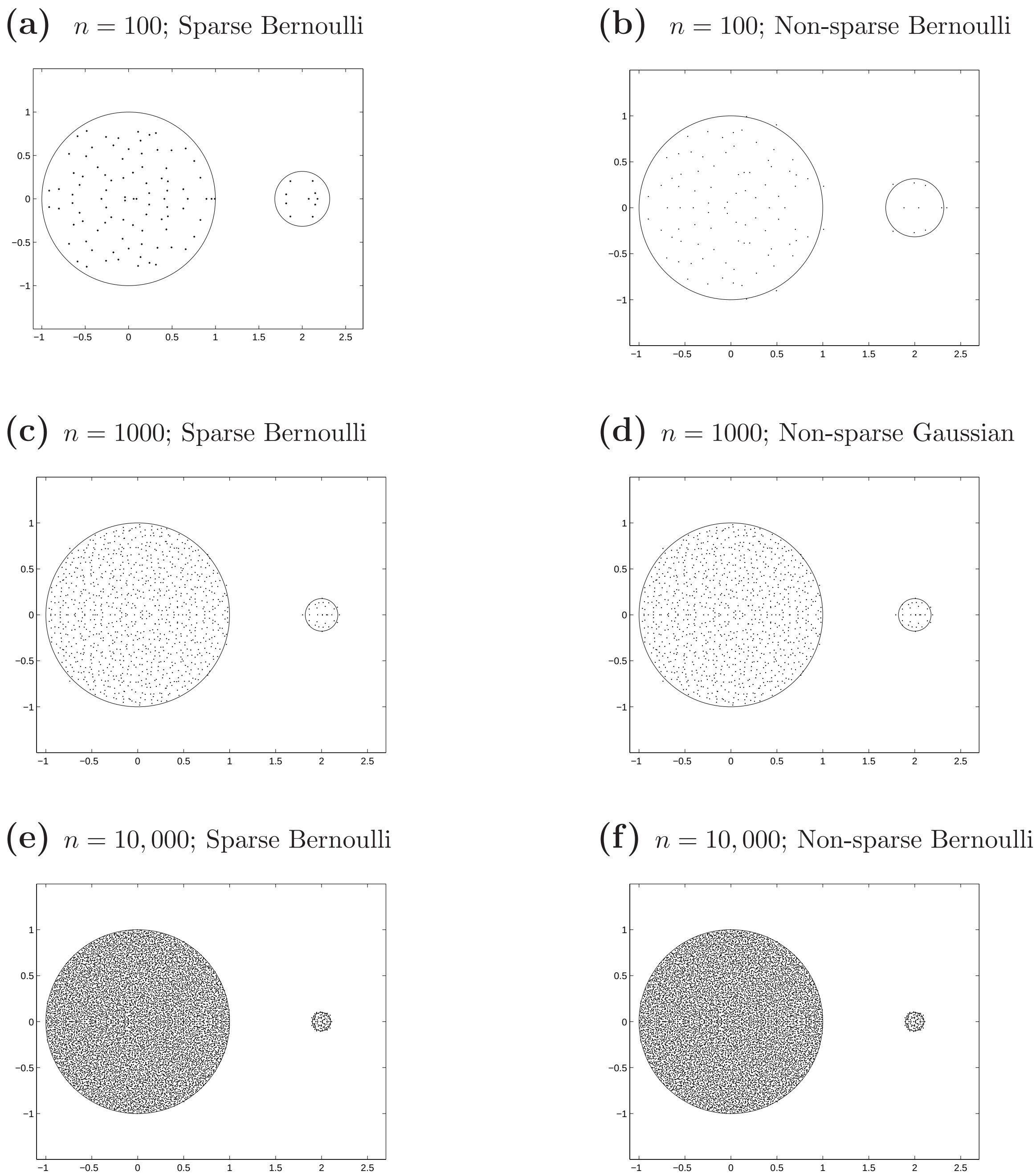


Figure 3: The six figures above illustrate that the circular law holds for for shifted sparse Bernoulli and shifted non-sparse Bernoulli random matrix ensembles. Each plot is of the eigenvalues of an n by n (with n as specified) random matrix of the form $M_n + X_n$, where M_n is a non-random diagonal matrix with the first $\lfloor\sqrt{n}\rfloor$ diagonal entries equal to $2\sqrt{n}$ and the remaining entries equal to zero, and X_n contains iid random entries. In the first column (figures (a), (c), and (e)), the matrices are sparse with parameter $\alpha = 0.4$, which means each entry is zero with probability $1 - \frac{1}{n^{0.6}}$, and in the second column (figures (b), (d), and (f)), the matrices are not sparse (i.e., $\alpha = 1$). The matrix ensembles are Bernoulli, so each non-zero entry is equally likely to be -1 or 1. As n increases, the ESDs in both the sparse and non-sparse cases approach the uniform distribution on the unit disk. Empirically, the small circle on the right, which has roughly $\sqrt{n}$ eigenvalues in and near it, shrinks until its contribution to the ESD is negligible (as drawn, the small circle has radius $n^{-1/4}$).

Remark 1.10. The proof of Lemma 1.9 illustrates that $\rho = 1/n$ is a transition point for sparse random variables of the type $\mathbb{I}_\rho \xi$ where the arguments for universality break down. Notably, the proof of Lemma 1.9 also works for α depending on n so long as $\alpha \log n$ tends to infinity as $n \to \infty$; for example $\alpha = \frac{1}{\log \log n}$ is suitable. It would be interesting to see if the universality principle extends to parameters α that tend slowly to zero as $n \to \infty$.

1.3 Further directions

There are a number of natural further directions to consider with respect to the sparse universality principle Theorem 1.5. One natural question is whether the condition that the real and complex parts are independent can be removed. The condition is not necessary in the non-sparse case and seems to be an artifact of the proof; more discussion is provided in Remark 2.4. Another natural question is whether Theorem 1.5 can be generalized to prove almost sure convergence in addition to proving convergence in probability. A result of Dozier and Silverstein [4] is one of the ingredients used in [27] to prove almost sure convergence; however, there does not seem to be a sparse analog of [4]. Proving a sparse analog of [4] would be a substantial step towards proving a universality principle with almost sure convergence (see Remark 2.5), though there may be other avenues as well. Finally, a general question of interest would be to study the rates of convergence for the universality principle. Convergence seems reasonably fast in the non-sparse case; however, empirical evidence indicates that convergence is slower in the sparse case and may in fact depend on the underlying type of random variables—see Figure 4 for an example. A bound on convergence rates in the non-sparse case where the $(2+\delta)$th moment is bounded is given in [25, Section 14].

1.4 Definitions of convergence and notation

Let X be a random variable taking values in a Hausdorff topological space. We say that X_n *converges in probability* to X if for every neighborhood N_X of X, we have

$$\lim_{n\to\infty} \Pr(X_n \in N_X) = 1.$$

Furthermore, we say that X_n *converges almost surely* to X if

$$\Pr(\lim_{n\to\infty} X_n = X) = 1.$$

If C_n is a sequence of random variables taking values in $\mathbb{R}$, we say that C_n is *bounded in probability* if

$$\lim_{K\to\infty} \liminf_{n\to\infty} \Pr(C_n \le K) = 1.$$

In the current paper, we are interested in how a randomly generated sequence of ESDs μ_{A_n} converges as $n \to \infty$, and so we will put the standard vague topology on the space of probability measures on $\mathbb{C}$. In particular, if μ_n and μ'_n are randomly generated sequences of measures on $\mathbb{C}$, then μ_n converges to μ'_n in probability if for every smooth function with compact support f and for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \Pr(\left| \int_{\mathbb{C}} f \, d\mu_n - \int_{\mathbb{C}} f \, d\mu'_n \right| \le \epsilon) = 1.$$

Furthermore, μ_n converges to μ'_n almost surely if for every smooth function with compact support f and for every $\epsilon > 0$, the expression $\left| \int_{\mathbb{C}} f \, d\mu_n - \int_{\mathbb{C}} f \, d\mu'_n \right|$ converges to 0 with probability 1.

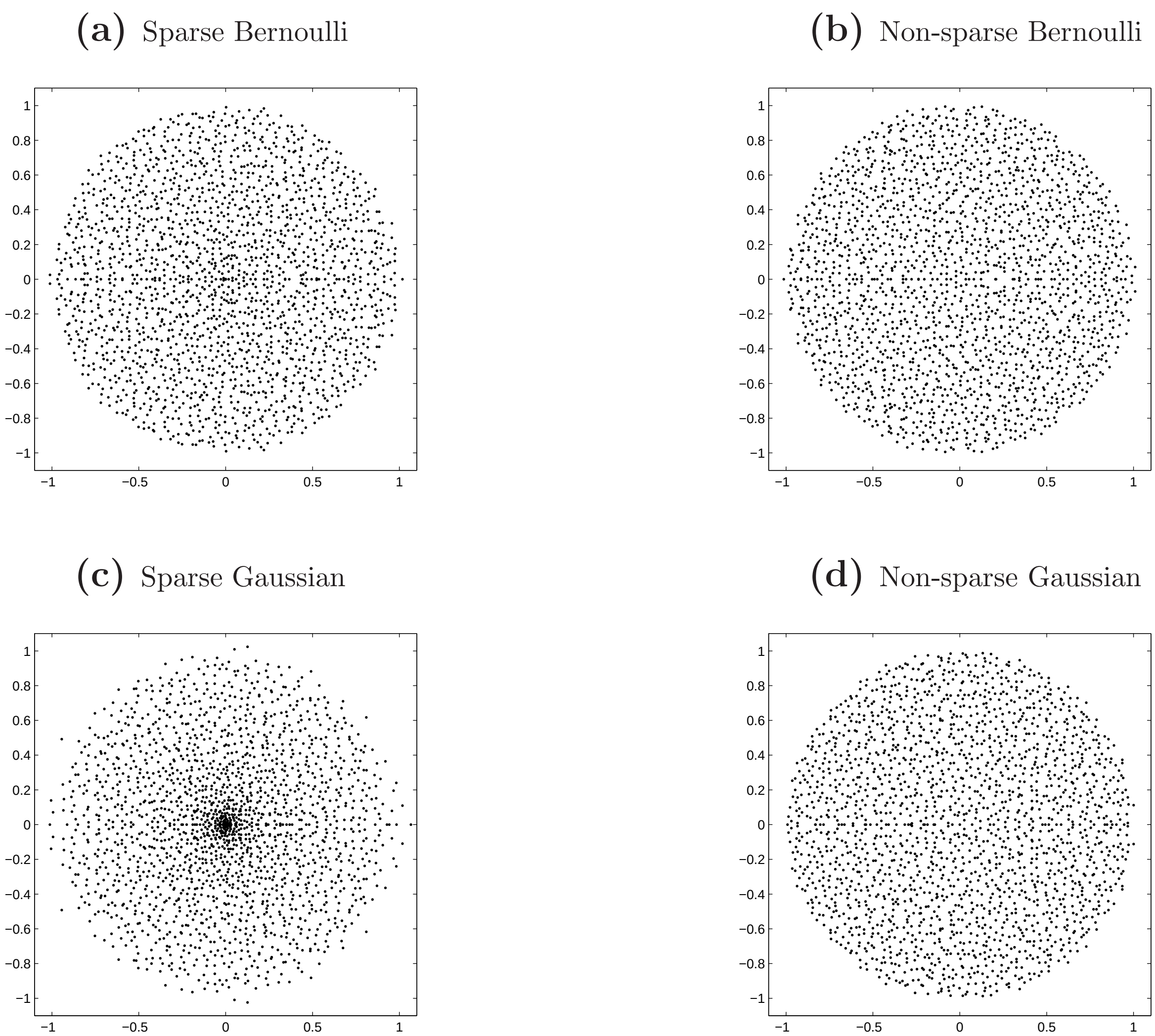


Figure 4: The four figures above indicate that the rates of convergence to the uniform distribution on the unit disk for for sparse Bernoulli and sparse Gaussian random matrix ensembles are apparently not the same as each other, and that in particular the sparse Gaussian case converges more slowly that the non-sparse case. Each plot is of the eigenvalues of a 2000 by 2000 random matrix with iid entries. In the first column (figures (a) and (c)), the matrices are sparse with parameter $\alpha = 0.2$, which means each entry is zero with probability $1 - \frac{1}{n^{0.8}}$, and in the second column (figures (b) and (d)), the matrices are not sparse (i.e., $\alpha = 1$). In the first row, both matrix ensembles are Bernoulli, so each non-zero entry is equally likely to be -1 or 1, and in the second row, the ensembles are Gaussian, so the non-zero entries are drawn from a Gaussian distribution with mean zero and variance one.

For functions f and g depending on n, we will make use of the asymptotic notation $f = O(g)$ to mean that there exists a positive constant c (independent of n) such that $f \leq cg$ for all sufficiently large n. Also, we will use the asymptotic notation $f = o(g)$ to mean that $f/g \to 0$ as $n \to \infty$.

1.5 Paper outline

Recall that the sparseness is determined by $\rho := n^{-1+\alpha}$. In the remaining sections, we will follow the approach used in [27] to prove a universality principle for sparse random matrices when $\alpha > 0$. In Section 2, we outline the main steps of the proof, highlighting a general result about convergence of ESDs from [27] that essentially reduces the question of convergences of ESDs to a question of convergence of the determinants of the corresponding matrices (one of which is sparse, and the other of which is not). Section 3 gives a proof of a sparse version of the necessary result on convergence of determinants based on a least singular value bound for sparse matrices in [25] and two lemmas, which are proved in Sections 4 and 5, respectively. In Section 5, we make use of a result of Chatterjee [3] which requires adapting Krishnapur's ideas in [27, Appendix C] to a sparse context ([27, Appendix C] is dedicated to proving a universality principle for non-sparse random matrices where the entries are not necessarily iid).

2 Proof of Theorem 1.5

The following result was proven by Tao and Vu [27, Theorem 2.1] and can be applied directly in proving Theorem 1.5. All logarithms in this paper are natural unless otherwise noted.

Theorem 2.1. [27] *Suppose for each n that $A_n, B_n \in \mathbb{M}_n(\mathbb{C})$ are ensembles of random matrices. Assume that*

(i) *The expression*

$$\frac{1}{n^2}\|A_n\|_2^2 + \frac{1}{n^2}\|B\|_2^2 \tag{3}$$

is bounded in probability.

(ii) *For almost all complex numbers z,*

$$\frac{1}{n}\log\left|\det(\frac{1}{\sqrt{n}}A_n - zI)\right| - \frac{1}{n}\log\left|\det(\frac{1}{\sqrt{n}}B_n - zI)\right|$$

converges in probability to zero. In particular, for each fixed z, these determinants are non-zero with probability $1 - o(1)$ for all n.

Then, $\mu_{\frac{1}{\sqrt{n}}A_n} - \mu_{\frac{1}{\sqrt{n}}B_n}$ converges in probability to zero.

Note that a stronger version of the above theorem appears in [27, Theorem 2.1] which additionally gives conditions under which $\mu_{\frac{1}{\sqrt{n}}A_n} - \mu_{\frac{1}{\sqrt{n}}B_n}$ converges almost surely to zero.

The lemma below is a sparse version of [27, Lemma 1.7].

Lemma 2.2. *Let M_n, A_n, and B_n be as in Theorem 1.5. Then $\frac{1}{n^2}\|A_n\|_2^2$ and $\int_{\mathbb{C}} |z|^2 \, d\mu_{\frac{1}{\sqrt{n}}A_n}(z)$ are bounded in probability, and the same statement holds with B_n replacing A_n.*

Proof. Our proof is the same as the proof [27, Lemma 1.7], except that we need to use a sparse version of the law of large numbers (see Lemma A.1). By the Weyl comparison inequality for second moment (see [27, Lemma A.2]), it suffices to prove that $\frac{1}{n^2}\|A_n\|_2^2$ is bounded in probability, and by the triangle inequality along with Inequality (2), it thus suffices to show that $\frac{1}{n^2}\|X_n\|_2^2$ is bounded in probability. By the sparse law of large numbers (see Lemma A.1) and the fact that $\mathbb{E}|x|^2 < \infty$, we see that $\frac{1}{n^2}\|X_n\|_2^2$ is bounded in probability. The statement with B_n replacing A_n is exactly [27, Lemma 1.7]. ☐

The proof of Theorem 1.5 is completed by combining Theorem 2.1 and Lemma 2.2 with the following proposition:

Proposition 2.3. *Let $0 < \alpha \leq 1$ be a constant and let x be a random variable with mean zero and variance one. Assume that the real and complex parts of x are independent; namely, that $\mathrm{Re}(x)$ is independent of $\mathrm{Im}(x)$. Let X_n be the sparse matrix ensemble for x with parameter α, and let Y_n be the n by n matrix having iid copies of x for each entry (in particular, Y_n is* not *sparse). For each n, let M_n be a deterministic n by n matrix satisfying Inequality (2) and let $A_n := M_n + X_n$ and let $B_n := M_n + Y_n$. Then, for every fixed $z \in \mathbb{C}$, we have that*

$$\frac{1}{n}\log\left|\det\left(\frac{1}{\sqrt{n}}A_n - zI\right)\right| - \frac{1}{n}\log\left|\det\left(\frac{1}{\sqrt{n}}B_n - zI\right)\right| \tag{4}$$

converges in probability to zero.

One useful property of the determinant is that it may be computed in a number of different ways. In particular, for a matrix M, we have

$$|\det(M)| = \prod_{i=1}^{n}|\lambda_i(M)| = \prod_{i=1}^{n}\sigma_i(M) = \prod_{i=1}^{n}\mathrm{dist}(R_i, \mathrm{Span}\{R_1, \ldots, R_{i-1}\}), \tag{5}$$

where $\lambda_i(M)$ and $\sigma_i(M)$ are the eigenvalues and singular values of M, respectively, and where R_i denotes the i-th row of M.

In the remainder of the current section, we will outline the program for proving Proposition 2.3 and describe the differences between our proof and the proof of [27, Proposition 2.2]. As in [27], we will prove Proposition 2.3 by writing the determinant as a product of distances between the i-th row of a matrix and the span of the first $i-1$ rows (thanks to Equation (5)). Proposition 2.3 can then be proven via three main steps:

1. A bound on the least singular value due to Tao and Vu [25] for sparse and non-sparse random matrices is used to take care of terms very high dimensional subspaces (i.e., span of more than $n - n^{1-\alpha/6}$ rows).

2. Talagrand's inequality is used, along with other ideas from [27] to take care of terms with high dimension (i.e., span of more than $(1-\delta)n$ rows) not already dealt with by the previous step. Some care must be taken in the sparse case with the constant α in the exponent in order to use Talagrand's inequality, which is where the $\alpha/6$ comes from in the previous step.

3. A result of Chatterjee [3] along with new ideas in [27] are used to take care of the remaining terms. Here, the sparse case differs substantially from the non-sparse case, in that we must

use Chatterjee's result [3] in place of a result due to Dozier and Silverstein [4] used in [27]. This step in general follows Krishnapur [27, Appendix C], who investigates a universality principle for non-sparse random matrices with not necessarily iid entries, since there Dozier and Silverstein's result [4] cannot be applied. Our use of Chatterjee's result [3] requires the hypothesis that the real and imaginary parts are independent, and this is the only place where that hypothesis is used.

Remark 2.4. It would be interesting to prove Theorem 1.5 without the hypothesis that $\mathrm{Re}(x)$ and $\mathrm{Im}(x)$ are independent. The only place that this hypothesis is used is when applying a result of Chatterjee [3, Theorem 1.1] (see Theorem 5.6), which was originally proven for independent real random variables. Two possible ways to proceed in removing the the hypothesis that $\mathrm{Re}(x)$ and $\mathrm{Im}(x)$ are independent would be proving a complex version of [3, Theorem 1.1] (though the result would require some re-phrasing, since the condition of differentiability is very different in the complex and real cases) or, alternatively, proving a sparse version of [4] (see Remark 2.5).

Remark 2.5. It would be natural to investigate a version of Theorem 1.5 where convergence in the almost sure sense is proved rather than convergence in probability. Typically, proving almost sure convergence is harder than proving convergence in probability, however, the universality principle in [27] is proven for both types of convergence, and so may provide a general approach to proving a universality principle for sparse random matrices with almost sure convergence. One of the steps in proving the universality principle of [27] in the almost sure sense uses a result due to Dozier and Silverstein [4]. In [4], a truncation argument is used that seems like it would need to be altered or replaced in order to prove a result for sparse random matrices. Another possible approach to proving a version of Theorem 1.5 for almost sure convergence would be to prove an analog of Chatterjee's [3, Theorem 1.1] (see Theorem 5.6) for almost sure convergence, though this might require a very different type of argument than the one used in [3]. A version of Lemma A.1 for almost sure convergence would also likely be necessary in any case.

3 Proof of Proposition 2.3

By shifting M_n by $zI\sqrt{n}$ (and noting that the new M_n still satisfies Inequality (2)), it is sufficient to prove that

$$\frac{1}{n}\log\left|\det(\frac{1}{\sqrt{n}}A_n)\right| - \frac{1}{n}\log\left|\det(\frac{1}{\sqrt{n}}B_n)\right|$$

converges to zero in probability.

Following the notation of [27], let $X_1, \ldots, X_n$ be the rows of A_n and let $Y_1, \ldots, Y_n$ be the rows of B_n. Let $Z_1, \ldots, Z_n$ denote the rows of M_n, and note that by Inequality (2) we have that

$$\sum_{j=1}^{n} \|Z_j\|_2^2 = O(n^2).$$

By re-ordering the rows of A_n, B_n, and M_n if necessary, we may assume that the rows $Z_{\lceil n/2\rceil}, \ldots, Z_n$ have the smallest norms, and so

$$\|Z_i\|_2 = O(\sqrt{n}), \text{ for } n/2 \le i \le n. \tag{6}$$

This fact will be used in part of the proof of Lemma 3.2.

For $1 \le i \le n$, let V_i be the $(i-1)$-dimensional space generated by $X_1, \dots, X_{i-1}$ and let W_i be the $(i-1)$-dimensional space generated by $Y_1, \dots, Y_{i-1}$. By standard formulas for the determinant (see Equation (5)), we have that

$$\frac{1}{n}\log\left|\det(\frac{1}{\sqrt{n}}A_n)\right| = \frac{1}{n}\sum_{i=1}^{n}\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i) \qquad \text{and}$$

$$\frac{1}{n}\log\left|\det(\frac{1}{\sqrt{n}}B_n)\right| = \frac{1}{n}\sum_{i=1}^{n}\log \operatorname{dist}(\frac{1}{\sqrt{n}}Y_i, W_i).$$

It is thus sufficient to show that

$$\frac{1}{n}\sum_{i=1}^{n}\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i) - \log \operatorname{dist}(\frac{1}{\sqrt{n}}Y_i, W_i) \tag{7}$$

converges in probability to zero. We will start by proving somewhat weak upper and lower bounds on $\operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i)$ and $\operatorname{dist}(\frac{1}{\sqrt{n}}Y_i, W_i)$ that hold for all i. For the upper bound, note that by Chebyshev's inequality we have $\Pr(\|X_i\|_2 > n^2) \le n^{-3}$, and thus by the Borel-Cantelli lemma, we have with probability 1 that $\|X_i\|_2 < n^2$ for all but finitely many n and for all i. This implies that, with probability 1,

$$\operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i) \le \|X_i\|_2 = n^{O(1)},$$

for all but finitely many n and for all i; and the same bound also holds for $\operatorname{dist}(\frac{1}{\sqrt{n}}Y_i, W_i)$. To show a lower bound, define $S_j^{(i)} := \operatorname{Span}(\{X_1, \dots, X_i\} \setminus \{X_j\})$ and define $A_n^{(i)}$ to be the i by n matrix consisting of the first i rows of A_n. By [27, Lemma A.4], we have

$$\sum_{j=1}^{i}\operatorname{dist}(X_j, S_j^{(i)})^{-2} = \sum_{j=1}^{i}\sigma_j(A_n^{(i)})^{-2},$$

and since $V_i = S_i^{(i)}$, we thus have the crude bound

$$\operatorname{dist}(X_i, V_i)^{-2} \le n\sigma_i(A_n^{(i)})^{-2}.$$

By Cauchy Interlacing (see [27, Lemma A.1]), we know that $\sigma_i(A_n^{(i)}) \ge \sigma_n(A_n)$, and thus we have

$$\frac{1}{n}\sigma_n(A_n) \le \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i),$$

and by the same reasoning,

$$\frac{1}{n}\sigma_n(B_n) \le \operatorname{dist}(\frac{1}{\sqrt{n}}Y_i, W_i).$$

Lower bounds on $\operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i)$ and $\operatorname{dist}(\frac{1}{\sqrt{n}}Y_i, W_i)$ will now follow from lower bounds on the least singular values of A_n and B_n which were proven in [25].

Lemma 3.1 (Least singular value bound for sparse random matrices)**.** [25] *Let $0 < \alpha \le 1$ be a constant and let x be a random variable with mean zero and variance one. Let X_n be the sparse matrix ensemble for x with parameter α, and let Y_n be the n by n matrix having iid copies of x for each entry (in particular, Y_n is* not *sparse). For each n, let M_n be a deterministic n by n matrix satisfying Inequality* (2) *and let $A_n := M_n + X_n$ and let $B_n := M_n + Y_n$. Then with probability 1 we have*

$$\sigma_n(A_n), \sigma_n(B_n) \ge n^{-O(1)}$$

for all but finitely many n.

Proof. Paraphrasing [27, proof of Lemma 4.1], the proof follows by combining [25, Theorem 2.5] (for the non-sparse matrix) and [25, Theorem 2.9] (for the sparse matrix) each with the Borel-Cantelli lemma, noting that the hypotheses of [25, Theorem 2.5] and [25, Theorem 2.9] are satisfied due to [25, Lemma 2.4] and Inequality (2). ☐

Thus, with probability 1 we have

$$\left|\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i)\right|, \left|\log \operatorname{dist}(\frac{1}{\sqrt{n}}Y_i, W_i)\right| \le O(\log n), \tag{8}$$

for all but finitely many n. In light of Inequality (8), the following two lemmas suffice to prove that the quantity in Display (7) converges in probability to zero.

Recall that α is the parameter used to determine the sparseness of the sparse matrix ensemble.

Lemma 3.2 (High-dimensional contribution)**.** *For every $\epsilon > 0$, there exists a constant $0 < \delta_\epsilon < 1/2$ such that for every $0 < \delta < \delta_\epsilon$ we have with probability 1 that*

$$\frac{1}{n} \sum_{(1-\delta)n \le i \le n - n^{1-\alpha/6}} \left|\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i)\right| = O(\epsilon)$$

for all but finitely many n.

Note that Lemma 3.2 with Y_i (which is not sparse) replacing X_i and with W_i replacing V_i was proven in [27, Lemma 4.2] with 0.99 replacing $1 - \alpha/6$. Alternatively, the non-sparse case follows from our proof of Lemma 3.2 if one sets $\alpha = 1$ (giving an exponent of $5/6$ in place of the exponent 0.99 used in [27, Lemma 4.2]). Also, note that for all sufficiently large n, we may assume that Equation (6) holds for all i relevant to Lemma 3.2 above.

Lemma 3.3 (Low-dimensional contribution)**.** *For every $\epsilon > 0$, there exists $0 < \delta < \epsilon$ such that with probability at least $1 - O(\epsilon)$ we have*

$$\left|\frac{1}{n} \sum_{1 \le i \le (1-\delta)n} \log\left(\operatorname{dist}\left(\frac{1}{\sqrt{n}}X_i, V_i\right)\right) - \log\left(\operatorname{dist}\left(\frac{1}{\sqrt{n}}Y_i, W_i\right)\right)\right| = O(\epsilon)$$

for all but finitely many n.

To complete the proof of Proposition 2.3, one may combine Lemma 3.2, [27, Lemma 4.2] (which is the non-sparse analog of Lemma 3.2), and Lemma 3.3. In particular, one may simply set ϵ in Lemma 3.3 equal to $\min\{\delta_1, \delta_2\}$, where δ_1 is the upper bound on δ from Lemma 3.2 and δ_2 is the corresponding upper bound on δ from [27, Lemma 4.2] (or from the non-sparse version of Lemma 3.2).

4 Proof of Lemma 3.2

Following [27], we will prove Lemma 3.2 in two parts, splitting the summands into cases where the log is positive and where the log is negative. The proof below follows the proof of [27, Lemma 4.2] closely, and we have included it in detail to make explicit the role of α, which determines the sparseness of the matrix A_n. One place where particular care must be taken with sparseness parameter α is in a truncation argument needed to apply Talagrand's inequality (see Subsection 4.3). There, we have made frequent use of the assumption that α is a positive constant, though it is possible that a very slowly decreasing α could also work—see Lemma 1.9 and Remark 1.10.

4.1 Positive log component

By the Borel-Cantelli lemma, the desired bound on the positive log component may be proven by showing

$$\sum_{n=1}^{\infty} \Pr\left(\frac{1}{n}\sum_{(1-\delta)n\le i\le n-n^{1-\alpha/6}} \max\{\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i), 0\} \ge \epsilon\right) < \infty. \tag{9}$$

We will use the crude bound $\max\{\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i), 0\} \le \max\{\log(\frac{\|X_i\|_2}{\sqrt{n}}), 0\}$. Note that if $2^{m_0} \le \frac{\|X_i\|_2}{\sqrt{n}} < 2^{m_0+1}$, then $m_0 \le \log_2(\frac{\|X_i\|_2}{\sqrt{n}}) < m_0 + 1$, and so

$$\sum_{m=0}^{\infty} \mathbb{1}_{\{\|X_i\|_2 \ge 2^m\sqrt{n}\}} = m_0 + 1 > \log_2(\frac{\|X_i\|_2}{\sqrt{n}}).$$

Thus,

$$\frac{1}{n}\sum_{(1-\delta)n\le i\le n-n^{1-\alpha/6}} \max\{\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i), 0\} \le \sum_{m=0}^{\infty}\frac{1}{n}\sum_{(1-\delta)n\le i\le n-n^{1-\alpha/6}} \mathbb{1}_{\{\|X_i\|_2\ge 2^m\sqrt{n}\}}. \tag{10}$$

If the left-hand side of Inequality (10) is at least ϵ for a given n, then we must have for some $m \ge 0$ that

$$\frac{1}{n}\sum_{(1-\delta)n\le i\le n-n^{1-\alpha/6}} \mathbb{1}_{\{\|X_i\|_2\ge 2^m\sqrt{n}\}} \ge \frac{2\epsilon}{(100+m)^2}. \tag{11}$$

We now have two cases to consider. For the first case, assume that the smallest m satisfying Inequality (11) satisfies $m \ge n^{1/5}$. Then for Inequality (11) to be satisfied, there exists some $1 \le i \le n$ such that $\|X_i\|_2 \ge 2^{n^{1/5}}\sqrt{n}$. By Chebyshev's inequality and Equation 6, we have that $\Pr(\|X_i\|_2 \ge 2^{n^{1/5}}\sqrt{n}) \le O(\frac{1}{2^{2n^{1/5}}})$, and thus the probability of such an i existing is at most $f(n) := 1-(1-c2^{-2n^{1/5}})^n$, where c is some constant. It is not hard to show that $f(n)n^2 \to 0$ as $n \to \infty$, and thus, for all sufficiently large n, we have the probability that there exists an i such that $\|X_i\|_2 \ge 2^{n^{1/5}}\sqrt{n}$ is at most ϵ/n^2. Since this probability is summable in n, we have proved Inequality (9) in the first case.

For the second case, assume that the smallest m satisfying Inequality (11) satisfies $0 \le m < n^{1/5}$. In this case we will use Hoeffding's Inequality.

Theorem 4.1 (Hoeffding's Inequality [11]). *Let $\beta_1, \ldots, \beta_k$ be independent random variables such that for $1 \le i \le k$ we have*

$$\Pr(\beta_i - \mathbb{E}(\beta_i) \in [0,1]) = 1.$$

Let $S := \sum_{i=1}^{k} \beta_i$. Then

$$\Pr(S \ge kt + \mathbb{E}(S)) \le \exp\left(-2kt^2\right).$$

The random variables β_i will be $\mathbb{1}_{\{\|X_i\|_2 \ge 2^m\sqrt{n}\}}$, and thus we need to control $\Pr(\|X_i\|_2 \ge 2^m\sqrt{n})$ in order to bound $\mathbb{E}(S)$. By Equation (6) and Chebyshev's inequality, we have that

$$\Pr(\|X_i\|_2 \ge 2^m\sqrt{n}) \le O(\frac{1}{2^{2m}}). \tag{12}$$

We will take $k = n - n^{1-\alpha/6} - (1-\delta)n$, so we have that $\lim_{n\to\infty} \frac{k}{n} = \delta$. Also, δ_ϵ sufficiently small so that $\delta_\epsilon < \frac{\epsilon}{20000C}$, where C is the implicit constant in Inequality (12). If we take $t = \frac{n}{k}\left(\frac{\epsilon}{(100+m)^2}\right)$, we can compute that

$$\frac{kt}{n} + \frac{1}{n}\mathbb{E}(S) \le \frac{\epsilon}{(100+m)^2} + \frac{2\delta_\epsilon C}{2^{2m}} \le \frac{2\epsilon}{(100+m)^2}$$

for all sufficiently large n (the second inequality follows by taking n sufficiently large so that $k/n \le 2\delta < 2\delta_\epsilon$). Thus, by Hoeffding's Inequality and taking n sufficiently large, we have

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{(1-\delta)n \le i \le n-n^{1-\alpha/6}} \mathbb{1}_{\{\|X_i\|_2 \ge 2^m\sqrt{n}\}} \ge \frac{2\epsilon}{(100+m)^2}\right) &\le \exp\left(\frac{-n\epsilon^2}{\delta(100+m)^4}\right) \\ &\le \max\left\{\exp\left(\frac{-n\epsilon^2}{\delta(200)^4}\right), \exp\left(\frac{-n^{1/5}\epsilon^2}{16\delta}\right)\right\},\end{aligned}$$

where the last inequality follows from our assumption in this second case that $0 \le m \le n^{1/5}$. Thus, we have shown for all sufficiently large n and any $0 \le \delta < \delta_\epsilon$ that

$$\Pr\left(\frac{1}{n}\sum_{(1-\delta)n \le i \le n-n^{1-\alpha/6}} \max\{\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i), 0\} \ge 0\right) \le \max\left\{\exp\left(\frac{-n\epsilon^2}{\delta(200)^4}\right), \exp\left(\frac{-n^{1/5}\epsilon^2}{16\delta}\right)\right\}.$$

Finally, we note that the bounds from the two cases sum to at most $\epsilon/n^2 + \max\left\{\exp\left(\frac{-n\epsilon^2}{\delta(200)^4}\right), \exp\left(\frac{-n^{1/5}\epsilon^2}{16\delta}\right)\right\}$, which is summable in n, thus completing the proof for the positive log component.

4.2 Negative log component

By the Borel-Cantelli Lemma, it suffices to show that

$$\sum_{n=1}^{\infty} \Pr\left(\frac{1}{n}\sum_{(1-\delta)n \le i \le n-n^{1-\alpha/6}} \max\{-\log \operatorname{dist}(\frac{1}{\sqrt{n}}X_i, V_i), 0\} \ge \epsilon\right) < \infty. \tag{13}$$

Following the approach in [27], our main tool is the following lemma.

Proposition 4.2. *Let $0 < \alpha \le 1$ be a constant, let $1 \le d \le n - n^{1-\alpha/6}$, let $0 < c < 1$ be a constant, and let W be a deterministic d-dimensional subspace of $\mathbb{C}^n$. Let X be a row of A_n. Then*

$$\Pr(\operatorname{dist}(X, W) \le c\sqrt{n-d}) = 6\exp(-n^{\alpha/2})$$

for all n sufficiently large with respect to c and α.

We will give the proof of Proposition 4.2 in Subsection 4.3. The proof of the negative log component of Lemma 3.2 can be completed by using Proposition 4.2 and following the proof of [27, Lemma 4.2], which we paraphrase below.

Taking $c = 1/2$ in Proposition 4.2 and conditioning on V_i, we have that for each $(1-\delta)n \le i \le n - n^{\alpha/6}$ that

$$\Pr\left(\operatorname{dist}\left(\frac{1}{\sqrt{n}}X_i, V_i\right) > \frac{\sqrt{n-i+1}}{2\sqrt{n}}\right) \ge 1 - O(\exp(n^{-\alpha/2})).$$

Thus, the probability that

$$\operatorname{dist}\left(\frac{1}{\sqrt{n}}X_i, V_i\right) > \frac{\sqrt{n-i+1}}{2\sqrt{n}} \tag{14}$$

simultaneously for all $(1-\delta)n \le i \le n - n^{\alpha/6}$ is at least $1 - O(n^{-10})$ (in fact, better bounds are possible, but this is sufficient).

Finally, choosing δ_ϵ sufficiently small so that $\frac{\delta_\epsilon}{2}\log\frac{4}{\delta_\epsilon} < \epsilon$, we can take the log of Inequality (14) and sum in i to get that the probability in the summand of Inequality (13) in at most $O(n^{-10})$, and this is summable in n, completing the proof of Inequality (13).

4.3 Proof of Proposition 4.2

Recall that X has coordinates $a_i = \frac{\mathbb{I}_\rho x_i}{\sqrt{\rho}} + m_i$, where m_i is a fixed element (it comes from the matrix M_n), x_i is a fixed, mean zero, variance 1 random variable (it does not change with n), and $\rho = n^{-1+\alpha}$ where $0 < \alpha \le 1$ is a constant. The proof of Proposition 4.2 closely follows the proof of [27, Proposition 5.1], and we give the details below to highlight how the proof must be modified to accommodate sparseness with parameter α. In particular, care must be taken with the value of α in the following three steps: first, when reducing to the case where the sparse random variables are bounded (since sparseness requires scaling by $1/n^{-1+\alpha}$), second, when showing that the sparse random variables restricted to the bounded case still have variance tending to 1 as $n \to \infty$, and third, when applying Talagrand's inequality where one must keep track of α in the exponent on the upper bound.

Proof. First we reduce to the case where X has mean 0. Let $v = \mathbb{E}(X)$. (Note that v is the row of M_n corresponding to X).

Note that $\operatorname{dist}(X, W) \ge \operatorname{dist}(X - v, \operatorname{Span}(W, v))$. Thus, by changing constants slightly (while still preserving $0 < c < 1$) and replacing d by $d+1$, it suffices to prove Proposition 4.2 in the mean zero case.

The second step is reducing to a case where the coordinates of X are bounded. In particular, we will show that, with probability at least $1 - 2\exp(-n^{\alpha/2})$, all but $n^{0.8}$ of the coordinates of X take values that are less than $n^{1/2-\alpha/4}$. Let $t_i := \mathbb{1}_{\{|a_i| \ge n^{(1-\alpha/2)/2}\}}$, and let $T := \sum_{i=1}^n t_i$. If $\mathbb{E}(T) = 0$, then with probability 1 we have that $|a_i| < n^{(1-\alpha/2)/2}$, and we are done with the reduction to the

case where the coordinates are bounded. Thus, it is left to show this reduction in the case where $\mathbb{E}(T) > 0$.

By Chernoff (see [24, Corollary 1.9]) we know that for every $\epsilon > 0$ we have

$$\Pr(|T - \mathbb{E}(T)| \geq \epsilon\mathbb{E}(T)) \leq 2\exp\left(-\min\left\{\frac{\epsilon^2}{4}, \frac{\epsilon}{2}\right\}\mathbb{E}(T)\right).$$

Since $\mathbb{E}(T) > 0$ by assumption, we may set $\epsilon := \frac{n^{0.8}}{\mathbb{E}(T)} - 1$. By Chebyshev's inequality, we have $\Pr(|a_i| \geq n^{(1-\alpha/2)/2}) \leq n^{-1+\alpha/2}$ for all $1 \leq i \leq n$, and thus $\mathbb{E}(T) = n\mathbb{E}(t_i) \leq n^{\alpha/2}$, which implies that $\epsilon \geq n^{0.8-\alpha/2} - 1 \geq 2$ for large n. Here we used the fact that $0 < \alpha/2 \leq .5$. Using the Chernoff bound we have

$$\begin{aligned}
\Pr(T \geq (1+\epsilon)\mathbb{E}(T) = n^{0.8}) &\leq 2\exp\left(-\frac{\epsilon}{2}\mathbb{E}(T)\right) \\
&\leq 2\exp\left(-n^{0.8}/2 + \mathbb{E}(T)/2\right) \\
&\leq 2\exp\left(-n^{0.8}/2 + n^{\alpha/2}/2\right) \\
&\leq 2\exp(-n^{0.8}/4) \\
&\leq 2\exp(-n^{\alpha/2}).
\end{aligned}$$

Thus, with probability at least $1 - 2\exp(-n^{\alpha})$, there are at most $n^{0.8}$ indices for which $|a_i| \geq n^{(1-\alpha/2)/2}$. For a subset $I \subset \{1, 2, \ldots, n\}$, let E_I denote the event that $I = \{i : |a_i| \geq n^{1/2-\alpha/4}; 1 \leq i \leq n\}$.

By the law of total probability, we have

$$\Pr(\mathrm{dist}(X, W) \leq c\sqrt{n-d}) \leq 2\exp(-n^{\alpha/2}) + \sum_{\substack{I \subset \{1,\ldots,n\} \\ |I| \leq n^{0.8}}} \Pr\left(\mathrm{dist}(X, W \leq c\sqrt{n-d}\middle| E_I\right)\Pr(E_I).$$

Thus, it is sufficient to show that

$$\Pr\left(\mathrm{dist}(X, W) \leq c\sqrt{n-d}\middle| E_I\right) \leq 4\exp(-n^{\alpha/2})$$

for each $I \subset \{1, \ldots, n\}$ such that $|I| \leq n^{0.8}$.

Fix such a set I. By renaming coordinates, we may assume that $I = \{n'+1, \ldots, n\}$ where $n - n^{0.8} \leq n' \leq n$. The next step is projecting away the coordinates in I. In particular, let $\pi : \mathbb{C}^n \to \mathbb{C}^{n'}$ be the orthogonal projection onto the first n' coordinates, and note that

$$\mathrm{dist}(X, W) \geq \mathrm{dist}(\pi(X), \pi(W)).$$

Thus, we can condition on $a_{n'+1}, \ldots, a_n$, adjust c slightly (without changing the fact that $0 < c < 1$), and (abusing notation to henceforth let n stand for n') see that it is sufficient to show

$$\Pr(\mathrm{dist}(X, W) \leq c\sqrt{n-d}|\, |a_i| < n^{1/2-\alpha/4}, \text{ for every } 1 \leq i \leq n) \leq 4\exp(-n^{\alpha/2}).$$

Lemma 4.3. *Let $\tilde{a}_i$ be the random variable a_i conditioned on $|a_i| < n^{1/2-\alpha/4}$. Then $\tilde{a}_i$ has variance $1 + o(1)$.*

Proof. By definition

$$\begin{aligned}
\operatorname{Var}(\tilde{a}_i) &= \mathbb{E}(|\tilde{a}_i|^2) - |\mathbb{E}(\tilde{a}_i)|^2 \\
&= \mathbb{E}(|a_i|^2 \,|\, |a_i| < n^{1/2-\alpha/4}) - \left|\mathbb{E}(a_i |\, |a_i| < n^{1/2-\alpha/4})\right|^2 \\
&= \frac{1}{\Pr(|a_i| < n^{1/2-\alpha/4})}\mathbb{E}(|a_i|^2 \, \mathbb{1}_{\{|a_i|<n^{1/2-\alpha/4}\}}) - \frac{1}{\Pr(|a_i| < n^{1/2-\alpha/4})^2}\left|\mathbb{E}(a_i \mathbb{1}_{\{|a_i|<n^{1/2-\alpha/4}\}})\right|^2.
\end{aligned}$$

Note that $a_i = \frac{\mathbb{I}_\rho x_i}{\sqrt{\rho}}$, and so $|a_i| < n^{1/2-\alpha/4}$ if and only if $|\mathbb{I}_\rho x_i| < n^{\alpha/4}$. Since x_i does not change with n, we see that $\Pr(|a_i| < n^{1/2-\alpha/4}) = \Pr(|\mathbb{I}_\rho x_i| < n^{\alpha/4}) \to 1$ as $n \to \infty$. Also, by Lemma 1.9, we know that $\mathbb{E}(|a_i|^2 \, \mathbb{1}_{\{|a_i|<n^{1/2-\alpha/4}\}}) \to \mathbb{E}(|a_i|^2) = 1$ and that $\mathbb{E}(a_i \mathbb{1}_{\{|a_i|<n^{1/2-\alpha/4}\}}) \to \mathbb{E}(a_i) = 0$. Thus, we have shown that $\tilde{a}_i$ has variance $1 + o(1)$. □

Next, we recenter $\tilde{a}_i$ by subtracting away its mean, and we call the result $\tilde{a}_i$. Note that this recentering does not change the variance. We will use the following version of Talagrand's inequality, quoted from [27, Theorem 5.2] (see also [14, Corollary 4.10]):

Theorem 4.4 (Talagrand's inequality)**.** *Let $\mathbf{D}$ be the unit disk $\{z \in \mathbb{C}, |z| \le 1\}$. For every product probability μ on $\mathbf{D}^n$, every convex 1-Lipschitz function $F : \mathbb{C}^n \to \mathbb{R}$, and every $r \ge 0$,*

$$\mu(|F - M(F)| \ge r) \le 4\exp(-r^2/8),$$

where $M(F)$ denotes the median of F.

Let $\tilde{X} = (\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n)$, and let μ be the distribution on $\mathbf{D}^n$ given by $\tilde{X}/2n^{1/2-\alpha/4}$. Let $F := \frac{1}{2}\operatorname{dist}\left(\frac{\tilde{X}}{2n^{1/2-\alpha/4}}, W\right)^2$, and note that F is convex and 1-Lipschitz, which follows since $\operatorname{dist}(\tilde{X}/2n^{1/2-\alpha/4}, W)$ is both convex and 1-Lipschitz (and also using the fact that $\operatorname{dist}(\tilde{X}/2n^{1/2-\alpha/4}, W) \le 1$, since $0 \in W$).

By Theorem 4.4 with $r = 3n^{\alpha/4}$, we have

$$\Pr\left(\left|\operatorname{dist}(\tilde{X}, W)^2 - M(\operatorname{dist}(\tilde{X}, W)^2)\right| \ge 12n^{\alpha/4}n^{1-\alpha/2}\right) \le 4\exp(-n^{\alpha/2}),$$

which implies that

$$\Pr(\operatorname{dist}(\tilde{X}, W)^2 \le M(\operatorname{dist}(\tilde{X}, W)^2) - 12n^{1-\alpha/4}) \le 4\exp(-n^{\alpha/2}). \tag{15}$$

Recall that $F = \frac{1}{2}\operatorname{dist}\left(\frac{\tilde{X}}{2n^{1/2-\alpha/4}}, W\right)^2$. Using Talagrand's inequality (Theorem 4.4) again, we will show that the mean of F is very close to the median of F. We compute

$$\begin{aligned}
|\mathbb{E}(F) - M(F)| \le \mathbb{E}\,|F - M(F)| &= \int_0^\infty \Pr(|F - M(F)| \ge t)\,dt \\
&\le \int_0^\infty 4\exp(-t^2/8)\,dt = 8\sqrt{2\pi}.
\end{aligned}$$

Thus, we have shown that

$$\left|\mathbb{E}(\operatorname{dist}(\tilde{X}, W)^2) - M(\operatorname{dist}(\tilde{X}, W)^2)\right| \le (32\sqrt{2\pi})n^{1-\alpha/2}. \tag{16}$$

Lemma 4.5. $\mathbb{E}(\mathrm{dist}(\tilde{X}, W)^2) = (1 + o(1))(n - d)$.

Proof. Let $\pi := (\pi_{ij})$ denote the orthogonal projection matrix to W. Note that $\mathrm{dist}(\tilde{X}, W)^2 = \sum_{i=1}^n \sum_{j=1}^n \tilde{a}_i \pi_{ij} \tilde{a}_j$. Since $\tilde{a}_i$ are iid, mean zero random variables, we have

$$\mathbb{E}(\mathrm{dist}(\tilde{X}, W)^2) = \mathbb{E}(|\tilde{a}_i|^2) \sum_{i=1}^n \pi_{ii} = \mathbb{E}(|\tilde{a}_i|^2)\,\mathrm{tr}(\pi).$$

The proof is completed by applying Lemma 4.3 and noting that the trace of π is $n - d$. $\square$

From Inequality (15), we see that it is sufficient to show that

$$M(\mathrm{dist}(\tilde{X}, W)^2) - 12n^{1-\alpha/4} \geq c^2(n - d).$$

Using Inequality (16) and Lemma 4.5 we have for sufficiently large n that

$$\begin{aligned} M(\mathrm{dist}(\tilde{X}, W)^2) - 12n^{1-\alpha/4} &\geq \mathbb{E}(\mathrm{dist}(\tilde{X}, W)^2) - (32\sqrt{2\pi})n^{1-\alpha/2} - 12n^{1-\alpha/4} \\ &\geq \left(c^2 + \frac{1-c^2}{2}\right)(n-d) - (32\sqrt{2\pi})n^{1-\alpha/2} - 12n^{1-\alpha/4} \\ &\geq c^2(n-d) + \left(\frac{1-c^2}{2}\right) n^{1-\alpha/6} - (32\sqrt{2\pi})n^{1-\alpha/2} - 12n^{1-\alpha/4} \\ &\geq c^2(n-d), \end{aligned}$$

where the last inequality follows from the fact that $\left(\frac{1-c^2}{2}\right) n^{1-\alpha/6} - (32\sqrt{2\pi})n^{1-\alpha/2} - 12n^{1-\alpha/4}$ is a positive quantity for sufficiently large n. Combining the above computation with Inequality 15 completes the proof of Proposition 4.2. $\square$

5 Proof of Lemma 3.3

Lemma 3.3 follows directly from the slightly more detailed statement in Lemma 5.1 given below. In this section, we will prove Lemma 5.1, closely follows the proof of [27, Lemma 4.3] with some changes. The biggest difference with the proof of [27, Lemma 4.3] is in the proof of Lemma 5.3, where we must adapt the approach of Krishnapur from [27, Appendix C] to a sparse setting (see Lemma 5.5). This is one critical juncture where it seems like it would take a new idea to prove almost sure convergence in place of convergence in probability. One possible approach would be proving a sparse version of [4] (which is used in [27] in the proof of almost sure convergence in the non-sparse case). Lemma 5.5 is also the only place where the assumption of independence between the real and complex parts is used, and it would be interesting to see if this assumption could be removed. Other notable differences from the proof of [27, Lemma 4.3] are that we must use Proposition 4.2 in place of [27, Proposition 5.1] and that we kill keep track of a lower bound on δ, which simplifies some steps in the proof.

Lemma 5.1. *For every $\epsilon_1 > 0$ and for all sufficiently small $\epsilon_2 > 0$, where ϵ_2 depends on ϵ_1 and other constants, the following holds. For every $\delta > 0$ satisfying*

$$\epsilon_2^2 < \delta \leq \frac{\epsilon_2}{40 \log(1/\epsilon_2)},$$

we have with probability $1 - O(\epsilon_1)$

$$\left| \frac{1}{n} \sum_{1 \le i \le (1-\delta)n} \log \operatorname{dist}\left(\frac{1}{\sqrt{n}} X_i, V_i \right) - \log \operatorname{dist}\left(\frac{1}{\sqrt{n}} Y_i, W_i \right) \right| = O(\epsilon_2)$$

for all but finitely many n*.*

As shown in [27, Section 6], it is sufficient to prove that with probability $1 - O(\epsilon_1)$ we have

$$\left| \frac{1}{n'} \sum_{i=1}^{n'} \log\left(\frac{1}{\sqrt{n}} \sigma_i(A_{n,n'}) \right) - \log\left(\frac{1}{\sqrt{n}} \sigma_i(B_{n,n'}) \right) \right| = O(\epsilon_2), \tag{17}$$

for all but finitely many n, where $n' = \lfloor (1-\delta)n \rfloor$, where $\sigma_i(A)$ denotes the i-th largest singular value of a matrix A, and where $A_{n,n'}$ denotes the matrix consisting of the first n' rows of A_n and $B_{n,n'}$ denotes the matrix consisting of the first n' rows of B_n.

Proving Equation (17) is equivalent to showing

$$\left| \int_0^\infty \log t \, d\nu_{n,n'}(t) \right| = O(\epsilon_2), \tag{18}$$

where $d\nu_{n,n'}$ is defined by the difference of the two relevant ESDs, namely:

$$d\nu_{n,n'} = d\mu_{\frac{1}{n'} A_{n,n'} A^*_{n,n'}} - d\mu_{\frac{1}{n'} B_{n,n'} B^*_{n,n'}}.$$

Following [27], we can prove Equation (18) by dividing the range of t into a few parts, which follows from Lemmas 5.2 (for large t), 5.3 (for intermediate-sized t), and 5.4 (for small t).

Lemma 5.2 (Region of large t)**.** *For every* $\epsilon_1 > 0$*, there exist constants* $\epsilon_2 > 0$ *and* R_{ϵ_2} *such that with probability* $1 - O(\epsilon_1)$ *we have*

$$\int_{R_{\epsilon_2}}^\infty |\log t| \left| d\nu_{n,n'}(t) \right| \le \epsilon_2.$$

Proof. By Lemma 2.2 and [27, Lemma A.2], we have that $\int_0^\infty t \left| d\nu_{n,n'}(t) \right|$ is bounded in probability. Thus, there exists a constant C_{ϵ_1} depending on ϵ_1 such that with probability $1 - O(\epsilon_1)$ we have

$$\int_0^\infty t \left| d\nu_{n,n'}(t) \right| \le C_{\epsilon_1}.$$

Choose $\epsilon_2 > 0$ sufficiently small with respect to ϵ_1 and C_{ϵ_1} so that

$$1 \ge 2 C_{\epsilon_1} \epsilon_2 \log\left(\frac{1}{\epsilon_2} \right).$$

Set $R_{\epsilon_2} = \left(\frac{1}{\epsilon_2} \right)^2$, and assume without loss of generality that $R_{\epsilon_2} > e$. Note that $\frac{t}{\log t}$ is increasing for $t \ge R_{\epsilon_2} > e$, and thus by the definition of ϵ_2 we have

$$\frac{C_{\epsilon_1}}{\epsilon_2} \log(t) \le t$$

whenever $t \geq R_{\epsilon_2}$. Thus, we have with probability $1 - O(\epsilon_1)$ that

$$\int_{R_{\epsilon_2}}^{\infty} |\log t| \left|d\nu_{n,n'}(t)\right| \leq \int_0^{\infty} \frac{\epsilon_2}{C_{\epsilon_1}} t \left|d\nu_{n,n'}(t)\right| \leq \epsilon_2.$$

$\square$

Lemma 5.3 (Region of intermediate t, namely $\epsilon_2^2 \leq t \leq R_{\epsilon_2}$)**.** *Define a smooth function $\psi(t)$ which equals 1 on the interval $[\epsilon_2^4, R_{\epsilon_2}]$, is supported on the interval $[\epsilon_2^4/2, 2R_{\epsilon_2}]$, is monotonically increasing on $(\epsilon_2^4/2, \epsilon_2^4)$, and is monotonically decreasing on $(R_{\epsilon_2}, 2R_{\epsilon_2})$.*

Then, with probability $1 - O(\epsilon_1)$ we have

$$\left|\int_0^{\infty} \psi(t) \log(t)\, d\nu_{n,n'}(t)\right| = O(\epsilon_2)$$

so long as $\delta \leq \frac{\epsilon_2}{40 \log(1/\epsilon_2)}$.

The main step in this proof is applying Lemma 5.5, whereas in the analogous step in the non-sparse case, [27] uses a result of Dozier and Silverstein [4], which proves almost sure convergence of the relevant distributions (rather than convergence in probability, which is the limit of Lemma 5.5). It would be interesting to see if a sparse analog of [4] is possible, especially since this might provide a way to remove the hypothesis requiring that the real and complex parts be independent, and it might further be one of the necessary components to proving a universality result for sparse random matrices with almost sure convergence instead of convergence in probability.

Proof. Using [27, Lemma A.1] and the upper bound on δ, it is possible to show that

$$\left|\int_0^{\infty} \psi(t) \log(t)\, d\nu_{n,n'}(t)\right| = \left|\int_0^{\infty} \psi(t) \log(t)\, d\nu_{n,n}(t)\right| + O(\epsilon_2)$$

(A possible alternative to the step above would be proving an analog of Lemma 5.5 for rectangular n by n' matrices.)

By Lemma 5.5 (see Subsection 5.1), we know that $d\nu_{n,n}$ converges in probability to zero, and thus

$$\left|\int_0^{\infty} \psi(t) \log(t)\, d\nu_{n,n}(t)\right| = O(\epsilon_2),$$

completing the proof. $\square$

The last step in proving Equation (18) and thus completing the proof of Lemma 5.1 is the following lemma:

Lemma 5.4 (Region of small t, namely $0 < t \leq \epsilon_2^4 < \delta^2$)**.** *With probability 1, we have*

$$\int_0^{\epsilon_2^4} |\log t| \left|d\nu_{n,n'}(t)\right| = O(\epsilon_2),$$

so long as $\delta \leq \frac{1}{2}\left(\frac{\epsilon_2}{\log(1/\epsilon_2)}\right)^{1/4}$.

Proof. The required upper bound on δ follows from the assumption that $\delta < \frac{\epsilon_2}{40 \log(1/\epsilon_2)}$. The proof is the same as the proof for [27, Lemma 6.6], with the small change that one must use Proposition 4.2 in place of [27, Proposition 5.1]. $\square$

5.1 Applying a theorem of Chatterjee

In this subsection, we follow the ideas used by Krishnapur in [27, Appendix C], where a central-limit-type theorem due to Chatterjee [3] was used to prove a universality result for random matrices with independent but not necessarily identically distributed entries. Lemma 5.5 below is analog of [27, Lemma C.3]. Recall that $\mathbb{I}_\rho$ is an iid copy of the random variable taking the value 1 with probability ρ and the value 0 with probability $1-\rho$, where $\rho = n^{-1+\alpha}$ where $0 < \alpha \le 1$ is a positive constant.

Lemma 5.5. *Let $\mathbf{X} = (X_{1,1}^{(0)}, X_{1,1}^{(1)}, X_{2,1}^{(0)}, X_{2,1}^{(1)}, \ldots)$ be an array of $2n^2$ independent real random variables, each of which is an iid copy of $X\mathbb{I}_\rho/\sqrt{\rho}$, where X is mean zero, variance 1, and let $\mathbf{Y} = (Y_{1,1}^{(0)}, Y_{1,1}^{(1)}, Y_{2,1}^{(0)}, Y_{2,1}^{(1)}, \ldots)$ be another array of $2n^2$ independent real random variables, each of which is a iid copy of a mean zero, variance 1 random variable Y. Let $A_n(\mathbf{X})$ denote the n by n random matrix having $X_{i,j}^{(0)} + \sqrt{-1}X_{i,j}^{(1)}$ for the (i,j) entry, and similarly for $A_n(\mathbf{Y})$. Let $\mu_{\frac{1}{n}A_n(\mathbf{X})A_n(\mathbf{X})^*}$ and $\mu_{\frac{1}{n}A_n(\mathbf{Y})A_n(\mathbf{Y})^*}$ denote the ESDs of $\frac{1}{n}A_n(\mathbf{X})A_n(\mathbf{X})^*$ and $\frac{1}{n}A_n(\mathbf{Y})A_n(\mathbf{Y})^*$, respectively. Then, $\mu_{\frac{1}{n}A_n(X)A_n(X)^*} - \mu_{\frac{1}{n}A_n(Y)A_n(Y)^*}$ converges in probability to zero as $n \to \infty$.*

Proof. Our approach will be applying [3, Theorem 1.1] in a similar way to [27, Lemma C.3].

Let $H_n(\mathbf{X}) := \begin{pmatrix} 0 & \frac{1}{\sqrt{n}}A_n(\mathbf{X})^* \\ \frac{1}{\sqrt{n}}A_n(\mathbf{X}) & 0 \end{pmatrix}$

Note that the eigenvalues of $H_n(\mathbf{X})$ with multiplicity are exactly the positive and negative square roots of the eigenvalues with multiplicity of $\frac{1}{n}A_n(\mathbf{X})A_n(\mathbf{X})^*$. Also, the same fact applies to $H_n(\mathbf{Y})$ and $\frac{1}{n}A_n(\mathbf{Y})A_n(\mathbf{Y})^*$. We will now follow the computation given in [3, Section 2.4]. It is sufficient to show that $\mu_{H_n(\mathbf{X})} - \mu_{H_n(\mathbf{Y})}$ converges in probability to zero as $n \to \infty$.

Let $u, v \in \mathbb{R}$ with $v \neq 0$ and let $z = u + \sqrt{-1}v$. Define a function $f : \mathbb{R}^{2n^2} \to \mathbb{C}$ by

$$f(\mathbf{x}) = \frac{1}{2n} \operatorname{tr}\left((H_n(\mathbf{x}) - zI)^{-1}\right).$$

Here $\mathbf{x} = (x_{i,j}^{(k)})_{1 \le i,j \le n; k \in \{0,1\}}$, where $x_{i,j}^{(0)}$ corresponds to the real part (namely, $X_{i,j}^{(0)}$ or $Y_{i,j}^{(0)}$) and $x_{i,j}^{(1)}$ corresponds to the complex part (namely, $X_{i,j}^{(1)}$ or $Y_{i,j}^{(1)}$).

Define $G : \mathbb{R}^{2n^n} \to \mathbb{C}^{(2n)^2}$ by

$$G(\mathbf{x}) = (H_n(\mathbf{x}) - zI)^{-1}.$$

All eigenvalues of $H_n(\mathbf{x})$ are real, and thus all eigenvalues of $H(\mathbf{x}) - zI$ are non-zero (since $v \neq 0$). Thus, $G(\mathbf{x})$ is well-defined. From the matrix inversion formula, each entry of $G(\mathbf{x})$ is a rational expression in $x_{i,j}^{(k)}$ for $1 \le i,j \le n$ and $k \in \{0,1\}$. Thus G is infinitely differentiable in each coordinate $x_{i,j}^{(k)}$.

In the remainder of this section, we will use the shorthand G for $G(\mathbf{x})$ and the shorthand H for $H_n(\mathbf{x})$.

Note that

$$\frac{\partial G}{\partial x_{i,j}^{(k)}} = -G \frac{\partial H}{\partial x_{i,j}^{(k)}} \tag{19}$$

(this can be seen by using the product rule and differentiating both sides of the equation $(H_n(\mathbf{x}) - zI)G = I$). The following three formulas follow from Equation (19) and the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any two square matrices A and B:

$$\begin{aligned}
\frac{\partial f}{\partial x_{i,j}^{(k)}} &= \frac{-1}{2n} \operatorname{tr}\left(\frac{\partial H}{\partial x_{i,j}^{(k)}} G^2 \right), \\
\frac{\partial^2 f}{(\partial x_{i,j}^{(k)})^2} &= \frac{1}{n} \operatorname{tr}\left(\frac{\partial H}{\partial x_{i,j}^{(k)}} G \frac{\partial H}{\partial x_{i,j}^{(k)}} G^2 \right), \qquad \text{and} \\
\frac{\partial^3 f}{(\partial x_{i,j}^{(k)})^3} &= \frac{-3}{n} \operatorname{tr}\left(\frac{\partial H}{\partial x_{i,j}^{(k)}} G \frac{\partial H}{\partial x_{i,j}^{(k)}} G \frac{\partial H}{\partial x_{i,j}^{(k)}} G^2 \right).
\end{aligned}$$

As in [3, Section 2.4], we will use the following facts to bound the partial derivatives of f. Note that $\mathrm{tr}(AB) = \|A\|_2 \|B\|_2$. Also, for A a k by k normal matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ and B any square matrix, we have $\max\{\|AB\|_2, \|BA\|_2\} \le (\max_{1 \le i \le k} \lambda_i) \|B\|_2$. By the definition of G, it is clear that the absolute value of the largest eigenvalue of G is at most $|v|^{-1}$. Also, by the definition of H, it is clear that $\frac{\partial H}{\partial x_{i,j}^{(k)}}$ is the matrix having $(\sqrt{-1})^k n^{-1/2}$ for the $(n+i, j)$ entry, having $(-\sqrt{-1})^k n^{-1/2}$ for the $(j, n+i)$ entry, and having zero for all other entries.

Thus, for all $1 \le i, j \le n$ and $k \in \{0, 1\}$, we have that

$$\begin{aligned}
\left| \frac{\partial f}{\partial x_{i,j}^{(k)}} \right| &\le \frac{1}{2n} \left\| \frac{\partial H}{\partial x_{i,j}^{(k)}} \right\| \|G^2\| \\
&\le \frac{1}{2n} \sqrt{\frac{2}{n}} \, |v|^{-1} \, \|I\| \\
&\le \frac{|v|^{-2}}{n}.
\end{aligned}$$

By similar means, we can compute

$$\begin{aligned}
\left| \frac{\partial^2 f}{(\partial x_{i,j}^{(k)})^2} \right| &\le \frac{1}{2n} \left\| \frac{\partial H}{\partial x_{i,j}^{(k)}} \right\| \left\| G \frac{\partial H}{\partial x_{i,j}^{(k)}} G^2 \right\| \\
&\le \frac{1}{n} \sqrt{\frac{2}{n}} \, |v|^{-3} \sqrt{\frac{2}{n}} \\
&\le \frac{2 \, |v|^{-3}}{n^2},
\end{aligned}$$

and

$$\begin{aligned}
\left| \frac{\partial^3 f}{(\partial x_{i,j}^{(k)})^3} \right| &\le \frac{3}{n} \left\| \frac{\partial H}{\partial x_{i,j}^{(k)}} \right\| \left\| G \frac{\partial H}{\partial x_{i,j}^{(k)}} G \frac{\partial H}{\partial x_{i,j}^{(k)}} G^2 \right\| \\
&\le \frac{6\sqrt{2} \, |v|^{-4}}{n^{5/2}}.
\end{aligned}$$

We will now apply the main theorem from [3]. First, we need the following definitions for a function $h : \mathbb{R}^N \to \mathbb{C}$. Let

$$\lambda_2(h) := \sup\left\{ \left| \frac{\partial h}{\partial x_{i,j}^{(k)}} \right|^2, \left| \frac{\partial^2 h}{(\partial x_{i,j}^{(k)})^2} \right| \right\}, \quad \text{and let}$$

$$\lambda_3(h) := \sup\left\{ \left| \frac{\partial h}{\partial x_{i,j}^{(k)}} \right|^3, \left| \frac{\partial^2 h}{(\partial x_{i,j}^{(k)})^2} \right|^{3/2}, \left| \frac{\partial^3 h}{(\partial x_{i,j}^{(k)})^3} \right| \right\}.$$

Theorem 5.6. [3] *Let $\mathbf{X} = (X_1, \dots, X_N)$ and $\mathbf{Y} = (Y_1, \dots, Y_N)$ be lists of independent, real-valued random variables such that $\mathbb{E}(X_i) = \mathbb{E}(Y_i)$ and $\mathbb{E}(X_i^2) = \mathbb{E}(Y_i^2)$ for all $1 \le i \le N$. Let $h : \mathbb{R}^N \to \mathbb{R}$ be thrice differentiable in each argument. If we set $U = h(\mathbf{X})$ and $V = h(\mathbf{Y})$, then for any thrice differentiable $g : \mathbb{R} \to \mathbb{R}$ and any $K > 0$,*

$$\begin{aligned} |\mathbb{E}g(U) - \mathbb{E}g(V)| \;\; &\le \;\; C_1(g)\lambda_2(h) \sum_{i=1}^{N} \big(\mathbb{E}(X_i^2; |X_i| > K) + \mathbb{E}(Y_i^2; |Y_i| > K)\big) \\ &\quad + C_2(g)\lambda_3(h) \sum_{i=1}^{N} \big(\mathbb{E}(|X_i|^3; |X_i| \le K) + \mathbb{E}(|Y_i|^3; |Y_i| \le K)\big) \end{aligned}$$

where $C_1(g) = \|g'\|_\infty + \|g''\|_\infty$ and $C_2(g) = \frac{1}{6}\|g'\|_\infty + \frac{1}{2}\|g''\|_\infty + \frac{1}{2}\|g'''\|_\infty$.

Thus, for our function f we have

$$\lambda_2(f) = \sup\left\{ \frac{|v|^{-4}}{n^2}, \frac{2\,|v|^{-3}}{n^2} \right\}, \quad \text{and}$$

$$\lambda_3(f) = \sup\left\{ \frac{|v|^{-6}}{n^3}, \frac{\sqrt{8}\,\left|v^{-9/2}\right|}{n^3}, \frac{6\sqrt{2}\,|v|^{-4}}{n^{5/2}} \right\}.$$

Theorem 5.6 requires h to be a real-valued function, thus we will apply Theorem 5.6 to $\mathrm{Re}(f)$ and $\mathrm{Im}(f)$ separately. Given $g : \mathbb{R} \to \mathbb{R}$ a thrice differentiable function, set $U = \mathrm{Re}(f(\mathbf{X}))$ and $V = \mathrm{Re}(f(\mathbf{Y}))$, where $\mathbf{X}$ and $\mathbf{Y}$ are as in the statement of Lemma 5.5 (notationally, set $N = 2n^2$ and define X_ℓ by $X_{1+2n(i-1)+2(j-1)+k} := X_{i,j}^{(k)}$). Noting that $\lambda_r(\mathrm{Re}\, f) \le \lambda_r(f)$, we may apply Theorem 5.6 to get

$$|\mathbb{E}g(U) - \mathbb{E}g(V)| \le \;\; C_1(g)\lambda_2(f) \sum_{k\in\{0,1\}} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}((X_{i,j}^{(k)})^2; \left|X_{i,j}^{(k)}\right| > K) + \mathbb{E}((Y_{i,j}^{(k)})^2; \left|Y_{i,j}^{(k)}\right| > K) \tag{20}$$

$$+C_2(g)\lambda_3(f) \sum_{k\in\{0,1\}} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}(\left|X_{i,j}^{(k)}\right|^3; \left|X_{i,j}^{(k)}\right| \le K) + \mathbb{E}(\left|Y_{i,j}^{(k)}\right|^3; \left|Y_{i,j}^{(k)}\right| \le K). \tag{21}$$

Choose $K = \epsilon\sqrt{n}$, where $\epsilon > 0$ is a small positive constant. The triple-sum term in Display (21) is bounded by ϵ times a constant depending only on g and v (here, we used that $\mathbb{E}(|X|^3; |X| \le K) \le$

$K\mathbb{E}(X^2)$ for any real random variable X). Also, the triple-sum term in Display (20) is bounded by another constant depending only on g and v times the quantity

$$\frac{1}{n^2}\sum_{k\in\{0,1\}}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}((X_{i,j}^{(k)})^2;\left|X_{i,j}^{(k)}\right|>\epsilon\sqrt{n})+\mathbb{E}((Y_{i,j}^{(k)})^2;\left|Y_{i,j}^{(k)}\right|>\epsilon\sqrt{n}).$$

Since the random variables $Y_{i,j}^{(k)}$ do not change with n, it is clear from monotone convergence that $\mathbb{E}((Y_{i,j}^{(k)})^2;\left|Y_{i,j}^{(k)}\right|>\epsilon\sqrt{n})\to 0$ as $n\to\infty$. Thus, it is sufficient to show that $\mathbb{E}((X_{i,j}^{(k)})^2;\left|X_{i,j}^{(k)}\right|>\epsilon\sqrt{n})\to 0$ as $n\to\infty$. Recall that $X_{i,j}^{(k)}$ is an iid copy of $X\mathbb{I}_\rho/\sqrt{\rho}$, where X is mean zero, variance 1. We have that

$$\mathbb{E}\left(\left|\frac{X\mathbb{I}_\rho}{\sqrt{\rho}}\right|^2;\left|\frac{X\mathbb{I}_\rho}{\sqrt{\rho}}\right|>\epsilon\sqrt{n}\right)\le\mathbb{E}\left(\left|\frac{X\mathbb{I}_\rho}{\sqrt{\rho}}\right|^2;|X|>\epsilon\sqrt{\rho n}\right)=\mathbb{E}\left(|X|^2;|X|>\epsilon\sqrt{\rho n}\right),$$

where the last equality follows by the independence of $\mathbb{I}_\rho$ and X. Finally, by monotone convergence again, we see that $\mathbb{E}\left(|X|^2;|X|>\epsilon\sqrt{\rho n}\right)\to 0$ as $n\to\infty$, completing the proof. $\square$

Acknowledgments

I would like to thank Van Vu for suggesting the project in the first place and for support throughout, and I would also like to thank Terence Tao for many useful discussions and ideas. I am also grateful for Persi Diaconis's advice on and dicussion about this project. This work was partly supported by an NSF Postdoctoral Research Fellowship.

A Sparse law of large numbers

Below we give a weak law of large numbers for sparse random variables of the sort considered in this paper. One of the ingredients to proving a universality result for the ESDs of sparse random matrices with almost sure convergence (rather than convergence in probability) would likely be a strong version of the sparse law of large numbers.

Lemma A.1 (Sparse Law of Large Numbers)**.** *Let ξ be a complex random variable such that $\mathbb{E}|\xi|<\infty$. Let X be a sparse version of ξ with parameter α, namely $X:=\mathbb{I}_\rho\xi/\rho$, where $\rho=n^{-1+\alpha}$, where $0<\alpha\le 1$ is a constant. Let m be a function of n such that $m=m(n)\ge n$. Then, if X_i is an iid copy of X for all $1\le i\le m$, we have for every $\epsilon>0$ that*

$$\lim_{n\to\infty}\Pr(\left|\frac{1}{m}\sum_{i=1}^{m}X_i-\mathbb{E}(\xi)\right|>\epsilon)=0.$$

I.e., $\frac{1}{m}\sum_{i=1}^{m}X_i$ converges to (ξ) in probability.

The proof below follows the general description given in [23].

Proof. We want to show for any small constant $\epsilon>0$ that $\frac{1}{m}\sum_{i=1}^{m}X_i=\mathbb{E}(\xi)+O(\epsilon)$ with probability at least $1-O(\epsilon)$.

By Lemma 1.9, we know that $\mathbb{E}(\left|\mathbb{1}_{\{|X|>n^{1-\alpha/2}\}}X\right|) \to 0$ as $n \to \infty$; and thus, we can choose n sufficiently large so that $\mathbb{E}(\left|\mathbb{1}_{\{|X|>n^{1-\alpha/2}\}}X\right|) \le \epsilon^2/4$ and $\left|\mathbb{E}(\mathbb{1}_{\{|X|\le n^{1-\alpha/2}\}}X) - \mathbb{E}(\xi)\right| \le \epsilon/4$.

Note that

$$\frac{1}{m}\sum_{i=1}^{m} X_i = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{|X_i|\le n^{1-\alpha/2}\}}X_i + \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{|X_i|>n^{1-\alpha/2}\}}X_i.$$

We thus have that

$$\Pr(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbb{E}(\xi)\right| > \epsilon) \le \Pr(\left|\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{|X_i|\le n^{1-\alpha/2}\}}X_i - \mathbb{E}(\xi)\right| > \epsilon/2) + \Pr(\left|\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{|X_i|>n^{1-\alpha/2}\}}X_i\right| > \epsilon/2). \tag{22}$$

Applying Chebyshev's inequality to the first term and using the fact that $\left|\mathbb{E}(\mathbb{1}_{\{|X|\le n^{1-\alpha/2}\}}X) - \mathbb{E}(\xi)\right| \le \epsilon/4$, we have that

$$\Pr(\left|\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{|X_i|\le n^{1-\alpha/2}\}}X_i - \mathbb{E}(\xi)\right| > \epsilon/2) \le \frac{16n^{1-\alpha/2}}{m\epsilon^2}\mathbb{E}\left|X\right|.$$

Applying Markov's inequality to the second term, we have

$$\Pr(\left|\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{|X_i|>n^{1-\alpha/2}\}}X_i\right| > \epsilon/2) < \epsilon/2.$$

Plugging these two estimates into Inequality (22) and using the facts that $\mathbb{E}(|X|) = \mathbb{E}(|\xi|)$ and $m \ge n$, we have that

$$\Pr(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mathbb{E}(\xi)\right| > \epsilon) \le \frac{16\mathbb{E}(|\xi|)}{\epsilon^2 n^{\alpha/2}} + \frac{\epsilon}{2},$$

which is less than ϵ for n chosen sufficiently large with respect to ϵ, α, and $\mathbb{E}(|\xi|)$. □

References

[1] Z. D. Bai. Circular law. *Ann. Probab.*, 25(1):494–529, 1997.

[2] Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010.

[3] Sourav Chatterjee. A simple invariance theorem. arXiv:math/0508213v1, pages 1–14, 12 Aug 2005.

[4] R. Brent Dozier and Jack W. Silverstein. On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *J. Multivariate Anal.*, 98(4):678–694, 2007.

[5] M. M. Duras, K. Sokalski, and P. Sułkowski. Statistical properties of sparse Gaussian random symmetrical ensemble. *Acta Phys. Polon. B*, 28(5):1023–1038, 1997.

[6] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9(4):543–560, 1988.

[7] V. L. Girko. The circular law. *Teor. Veroyatnost. i Primenen.*, 29(4):669–679, 1984.

[8] V. L. Girko. The strong circular law. Twenty years later. II. *Random Oper. Stochastic Equations*, 12(3):255–312, 2004.

[9] F. Götze and A. Tikhomirov. On the circular law. arXiv:math/0702386v1, pages 1–35, 13 Feb 2007.

[10] F. Götze and A. Tikhomirov. The circular law for random matrices. *Ann. Probab.*, 38(4):1444–1491, 2010.

[11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[12] Manjunath Krishnapur and Van Vu. in preparation.

[13] Reimer Kühn. Spectra of sparse random matrices. *J. Phys. A*, 41(29):295002, 21, 2008.

[14] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[15] M. L. Mehta. *Random matrices and the statistical theory of energy levels*. Academic Press, New York, 1967.

[16] Taro Nagao and Toshiyuki Tanaka. Spectral density of sparse sample covariance matrices. *J. Phys. A*, 40(19):4973–4987, 2007.

[17] Guangming Pan and Wang Zhou. Circular law, extreme singular values and potential theory. *J. Multivariate Anal.*, 101(3):645–656, 2010.

[18] G. J. Rodgers and A. J. Bray. Density of states of a sparse random matrix. *Phys. Rev. B (3)*, 37(7):3557–3562, 1988.

[19] Tim Rogers, Isaac Pérez Castillo, Reimer Kühn, and Koujin Takeda. Cavity approach to the spectral density of sparse symmetric random matrices. *Phys. Rev. E (3)*, 78(3):031116, 6, 2008.

[20] Shunsuke Sato and Kingo Kobayashi. Asymptotic distribution of eigenvalues and degeneration of sparse random matrices. *Bull. Math. Statist.*, 17(3–4):83–99, 1976/77.

[21] Guilhem Semerjian and Leticia F. Cugliandolo. Sparse random matrices: the eigenvalue spectrum revisited. *J. Phys. A*, 35(23):4837–4851, 2002.

[22] Sasha Sodin. The Tracy-Widom law for some sparse random matrices. *J. Stat. Phys.*, 136(5):834–841, 2009.

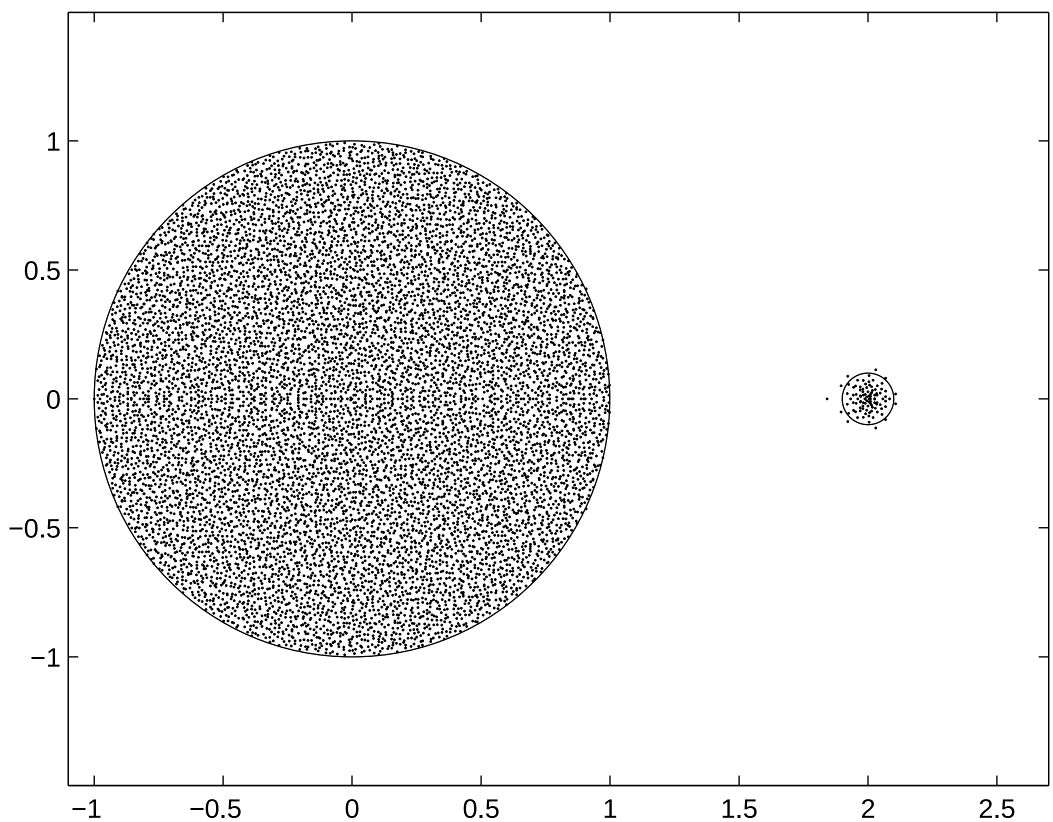
[23] Terence Tao. The strong law of large numbers. In the blog *What's new*. `http://terrytao.wordpress.com/`, 18 June 2008.

[24] Terence Tao and Van Vu. *Additive combinatorics*, volume 105 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2006.

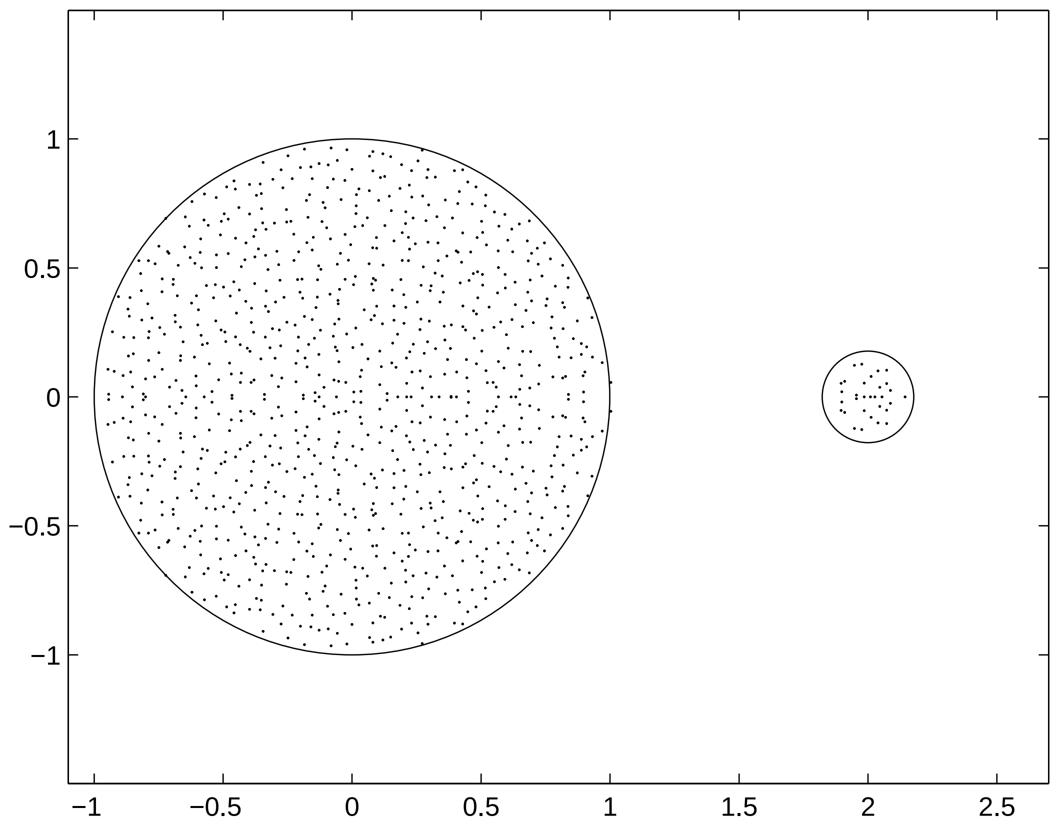
[25] Terence Tao and Van Vu. Random matrices: the circular law. *Commun. Contemp. Math.*, 10(2):261–307, 2008.

[26] Terence Tao and Van Vu. From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices. *Bull. Amer. Math. Soc. (N.S.)*, 46(3):377–396, 2009.

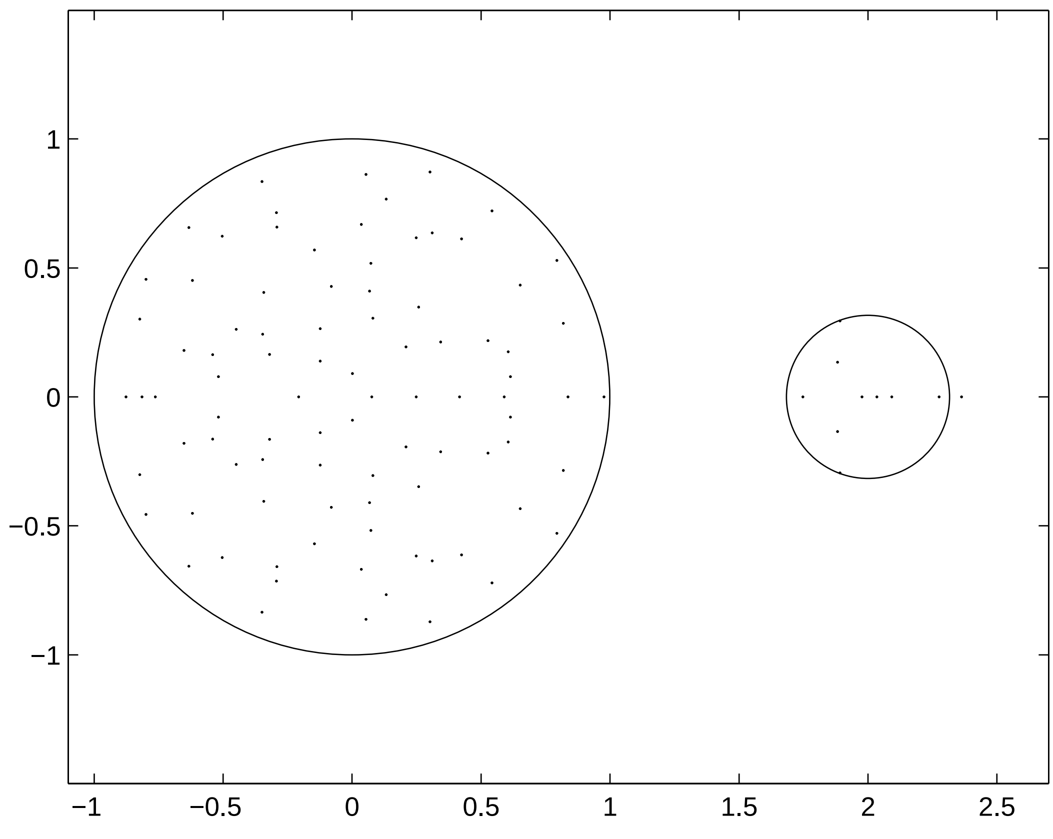
[27] Terence Tao, Van Vu, and Manjunath Krishnapur. Universality of ESDs and the circular law. *Ann. Probab.*, 38(5):2023–2065, 2010.

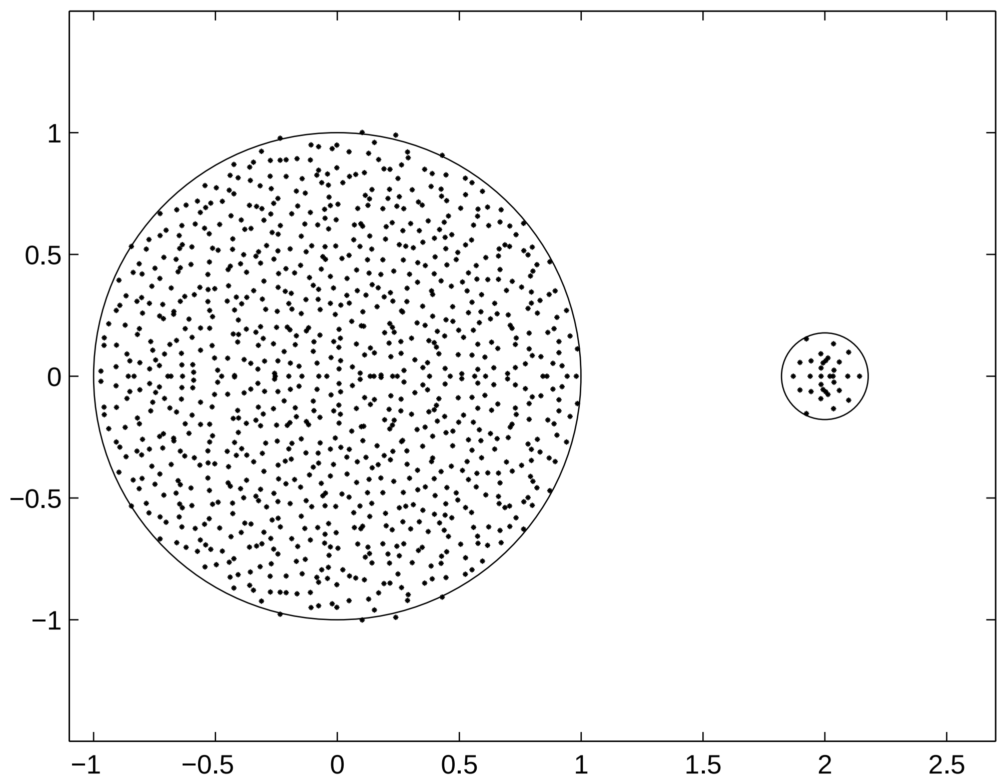


1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
1.5
2
2.5



1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
1.5
2
2.5


1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
1.5
2
2.5



1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1
1.5
2
2.5